A dynamic factor model approach to incorporate Big Data in state space models for official statistics

A dynamic factor model approach to incorporate Big Data in state space models for official statistics *

Caterina Schiavoni †[1,2], Franz Palm[2], Stephan Smeekes[2], and Jan van den Brakel[1,2]

[1]Statistics Netherlands, Heerlen
[2]Department of Quantitative Economics, Maastricht University

1st February 2019

Abstract

In this paper we consider estimation of unobserved components in state space models using a dynamic factor approach to incorporate auxiliary information from high-dimensional data sources. We apply the methodology to unemployment estimation as done by Statistics Netherlands, who uses a multivariate state space model to produce monthly figures for the unemployment using series observed with the labour force survey (LFS). We extend the model by including auxiliary series about job search behaviour from Google Trends and claimant counts, partially observed at higher frequencies. Our factor model allows for nowcasting the variable of interest, providing reliable unemployment estimates in real time before LFS data become available.

Keywords: high-dimensional data analysis, state space, factor models, nowcasting, unemployment, Google Trends.

1 Introduction

In this paper we investigate how "Big Data" can be incorporated into estimation of unobserved components using state space models. In particular, we investigate how auxiliary, noisy, data sources can be used to improve estimates of official statistics. Big Data sources have the problem that they are noisy and potentially (partly) irrelevant, and, as such, care must be taken when using them for the production of official statistics. We show that by using a dynamic factor model in state space form, relevant information can be extracted from such auxiliary high-dimensional data sources, while guarding against the inclusion of irrelevant data. We apply our methodology to the estimation of unemployment statistics.

Statistical information about a country's labour force is generally obtained from labour force surveys, since the required information is not available from registrations or other administrative data sources. The Dutch labour force survey (LFS) is based on a rotating panel design, where monthly household samples are

*This work was funded by the European Union under grant no. 07131.2017.003-2017.596. The views expressed in this paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. Previous versions of this paper have been presented at CFE-CM Statistics 2017, NESG 2018, SAE 2018, Methods for Big Data in Official Statistics, BigSurv 2018, the 29th $(EC)^2$ on Big Data Econometrics with Applications and at internal seminars organized by Maastricht University and Statistics Netherlands. We thank conference and seminar participants for their interesting comments. Additionally, we thank Marco Puts and Ole Mussmann for their help with the data collection. All remaining errors are our own.

†Corresponding author: Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. E-mail: c.schiavoni@maastrichtuniversity.nl.

observed five times with quarterly intervals. These figures are, however, considered too volatile to produce sufficiently reliable monthly estimates for the employed and the unemployed labour force at monthly frequency. For this reason Statistics Netherlands estimates monthly unemployment figures, together with its change, as unobserved components in a state space model where the observed series come from the monthly Dutch LFS, using a model originally proposed by Pfeffermann (1991). This method improves the precision of the monthly estimates for unemployment with sample information from previous periods, and can therefore be seen as a form of small area estimation (Rao and Molina, 2015). In addition it accounts for rotation group bias (Bailar, 1975), serial autocorrelation due to partial sample overlap and discontinuities due to several major survey redesigns (van den Brakel and Krieg, 2015).

Time series estimates for the unemployment can be further improved by including related auxiliary series. In this regard, Harvey and Chung (2000) propose a bivariate state space model to combine a univariate series of the monthly unemployed labour force derived from the UK LFS, with the univariate auxiliary series of claimant counts. The latter series represents the number of people claiming unemployment benefits. It is an administrative source, which is not available for every country, and, as for the Netherlands, it can be affected by the same publication delay of the labour force. In line with the aforementioned paper, we extend the state space model used by Statistics Netherlands in order to combine the survey data with a high-dimensional auxiliary series, as it could yield more information then a univariate one, which is not affected by publication lags and that can eventually be observed at a higher frequency than the labour force series.

This paper contributes to the existing literature by proposing a method to include a high-dimensional auxiliary series in a state space model in order to improve the (real-time) estimation of unobserved components. The model accounts for the rotating panel design underlying the sample survey series, combines series observed at different frequencies and deals with missing observations at the end of the sample due to publication delays. It handles the high dimensionality problem that arises from including a large number of series related to the unobserved components, by extracting their common factors. Moreover, we propose two extensions of the model, which allow the unobserved components of interest to depend on past values of the factors, and to model the cycle of the latter.

Besides claimant counts, the majority of the information related to unemployment is nowadays available on the internet; from job advertisements to resumé's templates and websites of recruitment agencies. We therefore follow the idea originating in Choi and Varian (2009), Askitas and Zimmermann (2009) and Suhoy (2009) of using job-related terms searched on Google in the Netherlands. Since 2004, these time series are freely downloadable in real-time from the Google Trends tool, on a monthly or higher frequency. As from the onset it is unclear which search terms are relevant, and if so, to which extent, care must be taken not to model spurious relationships with regards to the labour force series of interest, which could have a detrimental effect on the estimation of unemployment, such as happened for the widely publicized case of Google Flu Trends (Lazer et al., 2014).

Our method allows to exploit the high-frequency and/or real-time information of the auxiliary series, and to use it in order to nowcast the unemployment, before the publication of labour force data. As the number of search terms related to unemployment can easily become large, we employ the two-step estimator of Doz et al. (2011), which combines factor models with the Kalman filter, to deal both with the high-dimensionality of the auxiliary series, and with the estimation of the state space model. The above-mentioned estimator is generally used to improve the nowcast of variables that are observed, as the GDP (see Giannone et al. (2008) and Hindrayanto et al. (2016) for applications to the US and the euro area), which is not the case for the unemployment.

We evaluate the performance of our proposed method via Monte Carlo simulations and find that our method can yield large improvements in terms of MSFE of the unobserved components' estimators. We then assess whether the accuracy of the unemployment's estimation improves with our high-dimensional state space model, from both in-sample and out-of-sample results. The latter consists of a recursive nowcast. We

do not venture into forecasting exercises as Google Trends are considered to be more helpful in predicting the present rather than the future of economic activities (Choi and Varian, 2012). We conclude that Google Trends do not significantly improve the fit of the model, contrary to the claimant counts. Nonetheless, both empirical and simulation results show that our model is robust to the inclusion of a high-dimensional auxiliary series which does not have predictive power for the unobserved components of interest.

The remainder of the paper is organized as follows. Section 2.1 describes how data are collected with the Dutch LFS, and the state space model that is currently used by Statistics Netherlands to estimate the unemployment. Section 2.2 focuses on our proposed method to include a high-dimensional auxiliary series in the aforementioned model. Sections 3 and 4 report, respectively, the simulation and empirical results for our model. Section 5 concludes.

2 Methodology

2.1 The Dutch labour force model

The Dutch LFS is conducted as follows. Each month a stratified two-stage cluster design of addresses is selected. Strata are formed by geographical regions. Municipalities are considered as primary sampling units and addresses as secondary sampling units. All households residing on an address are included in the sample with a maximum of three (in the Netherlands there is generally one household per address). All household members with age of 16 or older are interviewed. Since October 1999, the LFS has been conducted as a rotating panel design. Each month a new sample, drawn according to the above-mentioned design, enters the panel and is interviewed five times at quarterly intervals. The sample that is interviewed for the j^{th} time is called the j^{th} wave of the panel, $j = 1, \ldots, 5$. After the fifth interview, the sample leaves the panel. This rotation design implies that in each month five samples are observed, which over time generate a five-dimensional time series of the unemployed labour force, defined as population total. Table 1 provides a visualization for the rotation panel design of the Dutch LFS.

quarter							
month							
A	BC	DEF	GHI	JKL	MNO	PQR	} wave 1
		ABC	DEF	GHI	JKL	MNO	} wave 2
			ABC	DEF	GHI	JKL	} wave 3
				ABC	DEF	GHI	} wave 4
					ABC	DEF	} wave 5

Table 1: Visualization for the rotation panel design of the Dutch LFS. Each capital letter represents a sample. Every month a new sample enters the panel and is interviewed five times at a quarterly frequency. After the fifth interview, the sample leaves the panel.

The remainder of this section describes the model that is currently used by Statistics Netherlands to estimate the Dutch unemployment (see van den Brakel and Krieg (2015) for more details). Let $y^k_{j,t}$ denote the generalized regression (GREG, i.e., design-based) estimate (Särndal et al., 1992) for the unemployment in month t based on the sample observed in wave j. Now $\boldsymbol{y}^k_t = (y^k_{1,t}, \ldots, y^k_{5,t})$ denotes the vector with the five GREG estimates for the uneployment in month t. The superscript $k > 1$ indicates that the vector is observed at the low frequency. We need this notation (introduced by Bańbura et al. (2013)) to distinguish between series observed at different frequencies, because later on we will make use of Google Trends about job search terms, which are available on a weekly basis. If $\boldsymbol{y}^k_t$ is observed at the monthly frequency, as in the case of the unemployed labour force, then $k = 4, 5$ if the high frequency series is observed at the weekly frequency, since a month can have either 4 or 5 weeks.

The unemployment is estimated, with the Kalman filter, as a state variable in a state space model where $\boldsymbol{y}_t^k$ represents the observed series. The measurement equation takes the form:

$$\boldsymbol{y}_t^k = \boldsymbol{\imath}_5 \theta_t^{k,y} + \begin{pmatrix} 0 \\ \boldsymbol{\lambda}_t^k \end{pmatrix} + \boldsymbol{e}_t^k. \tag{2.1}$$

where $\boldsymbol{\imath}_5$ is a 5-dimensional vector of ones, and $\theta_t^{k,y}$, i.e. the unemployment, is the common population parameter among the five-dimensional waves of the unemployed labour force. It is composed of a trend and a seasonal component:

$$\theta_t^{k,y} = L_t^{k,y} + S_t^{k,y}.$$

The transition equations for the level and the slope of the trend are, respectively:

$$\begin{aligned} L_t^{k,y} &= L_{t-1}^{k,y} + R_{t-1}^{k,y}, \\ R_t^{k,y} &= R_{t-1}^{k,y} + \eta_{R,t}^{k,y}, \quad \eta_{R,t}^{k,y} \sim N\left(0, \sigma_{R,y}^2\right), \end{aligned}$$

which characterize a smooth trend model, as the level does not have an innovation term. This implies that the level of the trend is $I(2)$, whereas the slope, i.e. the change in unemployment, is $I(1)$. The model originally contained an innovation term for the population parameter $\theta_t^{k,y}$. However, the maximum likelihood estimate for its variance tended to be zero and Bollineni-Balabay et al. (2017) showed via simulations that it is better to not include this term in the model.

The trigonometric stochastic seasonal component allows for the seasonality to vary over time, and it is modeled as in Durbin and Koopman (2012):

$$\begin{aligned} S_t^{k,y} &= \sum_{l=1}^{6} S_{l,t}^{k,y}, \\ \begin{pmatrix} S_{l,t}^{k,y} \\ S_{l,t}^{*k,y} \end{pmatrix} &= \begin{bmatrix} \cos(h_l) & \sin(h_l) \\ -\sin(h_l) & \cos(h_l) \end{bmatrix} \begin{pmatrix} S_{l,t-1}^{k,y} \\ S_{l,t-1}^{*k,y} \end{pmatrix} + \begin{pmatrix} \eta_{\omega,l,t}^{k,y} \\ \eta_{\omega,l,t}^{*k,y} \end{pmatrix}, \quad \begin{pmatrix} \eta_{\omega,l,t}^{k,y} \\ \eta_{\omega,l,t}^{*k,y} \end{pmatrix} \sim N\left(\mathbf{0}, \sigma_{\omega,y}^2 I_2\right), \end{aligned}$$

where $h_l = \frac{\pi l}{6}$, for $l = 1, \ldots, 6$.

Rotating panel designs generally suffer from Rotation Group Bias (RGB), which refers to the phenomena that there are systematic differences among the observations in the subsequent waves (Bailar, 1975). In the Dutch LFS the estimates for the unemployment based on the first wave are indeed systematically larger compared to the estimates based on the follow-up waves (van den Brakel and Krieg, 2015). This is the net results of different factors:

- Selective nonresponse among the subsequent waves, i.e., panel attrition;
- Systematic differences due to different data collection models that are applied to the waves;
- Differences in wording and questionnaire design used in the waves. In the first wave a block of questions is used to verify the status of the respondent on the labour force market. In the follow-up waves the questionnaire focuses on differences that occurred compared to the previous interview, instead of repeating the battery of questions;
- Panel conditioning effects, i.e., systematic changes in the behaviour of the respondents. For example, questions about activities to find a job in the first wave might increase the search activities of the unemployed respondents in the panel. Respondents might also systematically adjust their answers in the follow-up waves, since they learn how to keep the routing through the questionnaire as short as possible.

The answers from the first wave are hence assumed to be the most reliable ones and not to be affected by the RGB. Assuming $\lambda_{1,t}^k = 0$ implies that the Kalman filter estimates for $\theta_t^{k,y}$ in (2.1) are benchmarked to the level of the GREG series of the first wave. The four-dimensional state vector $\boldsymbol{\lambda}_t^k$ accounts for the RGB in the subsequent waves, as proposed in Pfeffermann (1991), and is modelled as a random walk because it aims at capturing time-dependent differences with respect to the first wave:

$$\begin{aligned}
&\lambda_{1,t}^k = 0, \\
&\lambda_{j,t}^k = \lambda_{j,t-1}^k + \eta_{\lambda,j,t}^k, \quad \eta_{\lambda,j,t}^k \sim N\left(0, \sigma_\lambda^2\right), \quad j = 2, \ldots, 5.
\end{aligned}$$

The rotating panel design also induces autocorrelation among the survey errors in the follow-up waves. In order to account for this autocorrelation, the survey errors are treated as state variables. Let $\boldsymbol{e}_t^k$ denote a five-dimensional vector containing the survey errors of the five waves, which follows the transition equation below.

$$\begin{aligned}
&e_{j,t}^k = c_{j,t}\tilde{e}_{j,t}^k, \quad c_{j,t} = \sqrt{\widehat{\text{var}}\left(y_{j,t}^k\right)}, \quad j = 1, \ldots, 5, \\
&\tilde{e}_{1,t}^k \sim N\left(0, \sigma_{\nu_1}^2\right), \\
&\tilde{e}_{j,t}^k = \delta \tilde{e}_{j-1,t-3}^k + \nu_{j,t}^k, \quad \nu_{j,t}^k \sim N\left(0, \sigma_{\nu_j}^2\right), \quad j = 2, \ldots, 5, \quad |\delta| < 1. \\
&\text{var}\left(\tilde{e}_{j,t}^k\right) = \sigma_{\nu_j}^2 / \left(1 - \delta^2\right), \quad j = 2, \ldots, 5.
\end{aligned}$$

The scaled sampling errors $\tilde{e}_{j,t}^k$, $j = 1, \ldots, 5$, account for the serial autocorrelation induced by the sampling overlap of the rotating panel. Samples in the first wave are observed for the first time and therefore its survey errors are not autocorrelated with survey errors of previous periods. The survey errors of the second to fifth wave are correlated with the survey errors of the previous wave three months before. Based on the approach proposed by Pfeffermann et al. (1998), van den Brakel and Krieg (2009) motivate that these survey errors should be modelled as an AR(3) process, without including the first two lags. Moreover, the survey errors of all waves are assumed to be proportional to the standard error of the GREG estimates. In this way the model accounts for heterogeneity in the variances of the survey errors, which are caused by changing sample sizes over time.

The structural time series model (2.1) as well as the models proposed in the following sections are fitted with the Kalman filter after putting the model in state space form. Initial values for the non-stationary state variables in the Kalman filter are obtained with a diffuse initialization. The state variables for the sampling errors are stationary and their initial values are obtained with an exact initialization. All the hyperparameters of the model are estimated by maximum likelihood using the L-BFGS-B optimization algorithm. The additional uncertainty of using maximum likelihood estimates for the hyperprarameters in the Kalman filter is ignored in the standard errors of the filtered state variables. Since the observed time series contains 168 periods, this additional uncertainty can be ignored. See also Bollineni-Balabay et al. (2017) for details. Both the simulation and estimation results in Sections 3 and 4 are obtained using the statistical software R.

2.2 High-dimensional auxiliary series

The Dutch labour force is subject to a one-month publication delay. In order to have more timely and precise estimates of the unemployment, we extend the model by including, respectively, auxiliary series about job search behaviour from weekly/monthly Google Trends and monthly claimant counts in the Netherlands.

Google Trends are indexes of search activity. Each index measures the fraction of queries that include the term in question in the chosen geography at a particular time, relative to the total number of queries

at that time. The maximum value of the index is set to be 100. According to the length of the selected period, the data can be downloaded at either monthly, weekly, or higher frequencies. The series are standardized according to the chosen period and their values can therefore vary according to the period's length (Stephens-Davidowitz and Varian, 2015). We use weekly and monthly Google Trends for each search term, and throughout the paper we denote them, respectively, with $\boldsymbol{x}_t^{GT}$ and $\boldsymbol{x}_t^{k,GT}$.

Figure 1 displays the time series of the five waves of the unemployed labour force, together with the claimant counts and an example of job-related Google query. They all seem to be following the same trend, which already shows the potential of using this auxiliary information in estimating the unemployment.

We denote the dimensionality of the vector $\boldsymbol{x}_t^{GT}$ by n, which can be large. Moreover, not all job search terms might be relevant in explaining the unemployment. We therefore need to address the high-dimensionality problem of these auxiliary series and make our model not too dependent on the individual Google Trends. Factor models serve this purpose by retaining the information of these time series in few common factors.

Moreover, when dealing with mixed frequency variables and with publication delays, we can encounter "jagged edge" datasets, which have missing values at the end of the sample period. The Kalman filter computes a prediction for the unobserved components in presence of missing observations for the respective observable variables.

The two-step estimator by Doz et al. (2011) combines factor models with the Kalman filter and hence addresses both of these issues. In the remainder of this section we explain how this estimator can be employed to nowcast the lower-frequency variable using information from the higher-frequency or real time variables.

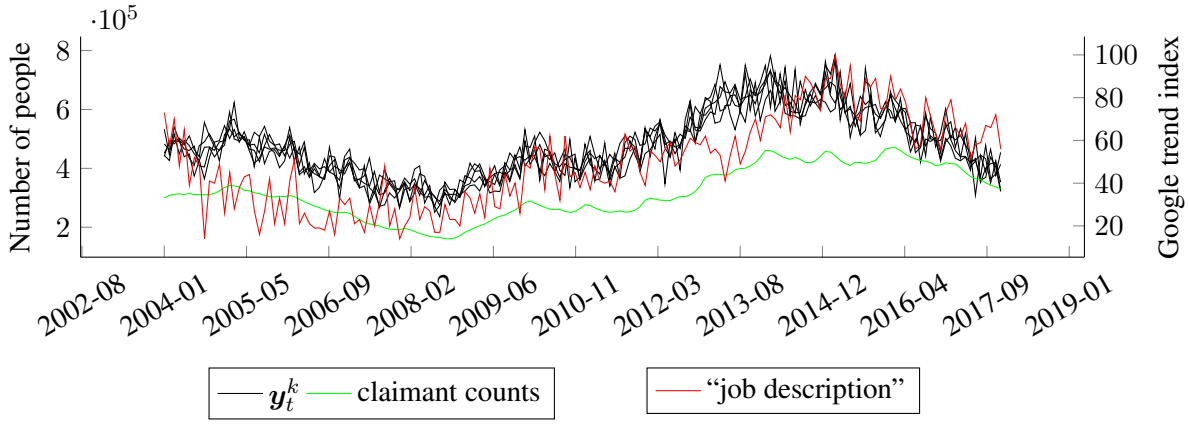

Figure 1: Monthly time series of the five waves of the Dutch unemployed labour force $\boldsymbol{y}_t^k$, the claimant counts and the Google search term "job description" in the Netherlands. The period starts in January 2004 and ends in December 2017.

2.2.1 Two-step estimator

We consider the state space representation of the dynamic factor model, with respective measurement and transition equations, as we would like to link it to the state space model used to estimate the unemployment (equation (2.1)):

$$\begin{aligned} \boldsymbol{x}_t^{GT} &= \Lambda \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t, \\ \boldsymbol{f}_t &= \boldsymbol{f}_{t-1} + \boldsymbol{u}_t. \end{aligned} \tag{2.2}$$

Notice that we are making the assumption of $\boldsymbol{x}_t^{GT}$ being $I(1)$ of dimension n, and $\boldsymbol{f}_t$ being $I(1)$ of dimension r. In section 2.2.2 the need of this assumption will become clearer as we will explain how to

make use of the two step estimator in the labour force model. The intuition behind it is that the factors and the change in unemployment, $R_t^{k,y}$, must have the same order of integration.

Bai (2004) proves the consistency of the estimator of $I(1)$ factors by principal component analysis (PCA), under usual assumptions as limited time and cross-sectional dependence and stationarity of the idiosyncratic components $\boldsymbol{\varepsilon}_t$, and non-trivial contributions of the factors to the variance of $\boldsymbol{x}_t$. We assume no cointegrating relationships among the factors. We further define the covariance matrix of the idiosyncratic components as $\mathrm{E}\left(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\right) = \Psi$, and of the innovations of the factors as $\mathrm{E}\left(\boldsymbol{u}_t\boldsymbol{u}_t'\right) = I_r$, for identifiability reasons.

The consistency of the two-step estimator has been originally proven in the stationary framework by Doz et al. (2011), and extended to the nonstationary case by Barigozzi and Luciani (2017).

The steps for the estimation proceed as follows:

1. The factor loadings Λ, the factors $\boldsymbol{f}_t$ and the covariance matrix of the idiosyncratic components Ψ are estimated by PCA (as in Bai (2004)) applied to a balanced dataset, meaning that in this first step the estimation is carried out without considering the missing observations at the end of the sample period.

2. $\hat{\Lambda}$ and $\hat{\Psi} = \mathrm{diag}\left(\hat{\psi}_{11}, \ldots, \hat{\psi}_{nn}\right)$ obtained in the previous step are kept fixed, and $\boldsymbol{f}_t$ are re-estimated with the Kalman filter applied to the approximated model on the entire dataset (i.e., including the missing observations at the end of the sample period). We hence need to condition on the information set Ω_υ to obtain $\hat{\boldsymbol{f}}_{t|\upsilon} = \mathrm{E}\left[\boldsymbol{f}_t|\Omega_\upsilon; \mathcal{M}_{\left(\hat{\Lambda},\hat{\Psi}\right)}\right]$, where $\mathcal{M}_{\left(\hat{\Lambda},\hat{\Psi}\right)}$ denotes the estimated model and υ is the time of a particular data release, which does not necessarily coincide with t (for instance in presence of publication delays). Nonetheless, in this paper it is assumed that $\upsilon = t$, i.e., all $\boldsymbol{x}_t^{GT}$ are observed at the same frequency and released at the same time without publication delays, which is the case for the Google Trends. The estimate of Λ is used in the state space model since its knowledge is needed in order to apply the Kalman filter. We fix $\hat{\Psi}$ because its high-dimensionality and associated curse of dimensionality complicates re-estimation by maximum likelihood. Restricting the covariance matrix of the idiosyncratic components as being diagonal is standard in the literature.

In formula (2.2) we do not make use of the superscript k, meaning that the two-step estimation is performed on the high frequency (weekly in our empirical case) variables. It is common practice in the literature, as explained in Giannone et al. (2008), to temporally aggregate the estimated factors to the low frequency of the observed macroeconomic variable of interest, and use them as regressors to nowcast the latter variable. As in our case the target variable itself, the unemployment, is unobserved, we have to nowcast it directly in the state space model used to estimate it.

2.2.2 Nowcasting in a high-dimensional state space model

In order to make use of the auxiliary series, we stack together the measurement equations for $\boldsymbol{y}_t^k$ and $\boldsymbol{x}_t^{k,GT}$, respectively (2.1) and the first equation of (2.2), and express them at the lowest frequency (in our case the monthly observation's frequency of $\boldsymbol{y}_t^k$). The transition equations for the RGB and survey error component in combination with the rotation scheme applied in the Dutch LFS hamper a formulation of the model on the high frequency. This means that $\boldsymbol{x}_t^{GT}$ needs to be first temporally aggregated from the high to the low frequency after the first step (which estimates Λ and Ψ). Since in practice $\boldsymbol{x}_t^{GT}$ are the $I(1)$ weekly Google Trends, which are flow variables as they measure the number of queries made during each week, they are aggregated according to the following rule (Bańbura et al., 2013):

$$\boldsymbol{x}_t^{k,GT} = \sum_{j=0}^{k-1} \boldsymbol{x}_{t-j}^{GT}, \quad t = k, 2k, \ldots. \tag{2.3}$$

The aggregated $\boldsymbol{x}_t^{k,GT}$ are then rescaled in order to be bounded again between 0 and 100. Equivalently, the temporal aggregation can be done by taking the average over the weeks:

$$\boldsymbol{x}_t^{k,GT} = \frac{1}{k}\sum_{j=0}^{k-1} \boldsymbol{x}_{t-j}^{GT}, \quad t = k, 2k, \ldots, \tag{2.4}$$

without additional rescaling. Harvey (1990) mentions that the aggregation according to equation (2.4) is suited for time-averaged stock variables, which are handled as flow variables from a statistical point of view.

In order to get the final model, we also include a measurement equation for the univariate auxiliary series of the claimant counts, assuming that its state vector, $\theta_t^{k,CC}$, has the same composition of our population parameter $\theta_t^{k,y}$ (i.e., composed of a smooth trend and a seasonal component):

$$\begin{pmatrix} \boldsymbol{y}_t^k \\ x_t^{k,CC} \\ \boldsymbol{x}_t^{k,GT} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\iota}_5 \theta_t^{k,y} \\ \theta_{t|\upsilon}^{k,CC} \\ \hat{\Lambda} \boldsymbol{f}_t^k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\lambda}_t^k \\ 0 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{e}_t^k \\ \varepsilon_t^{k,CC} \\ \boldsymbol{\varepsilon}_t^{k,GT} \end{pmatrix}, \quad \varepsilon_t^{k,CC} \sim N\left(0, \sigma_{\varepsilon,CC}^2\right), \tag{2.5}$$

$$\begin{pmatrix} \theta_t^{k,y} \\ \theta_t^{k,CC} \end{pmatrix} = \begin{pmatrix} L_t^{k,y} \\ L_t^{k,CC} \end{pmatrix} + \begin{pmatrix} S_t^{k,y} \\ S_t^{k,CC} \end{pmatrix}, \tag{2.6}$$

$$\begin{pmatrix} L_t^{k,y} \\ L_t^{k,CC} \end{pmatrix} = \begin{pmatrix} L_{t-1}^{k,y} \\ L_{t-1}^{k,CC} \end{pmatrix} + \begin{pmatrix} R_{t-1}^{k,y} \\ R_{t-1}^{k,CC} \end{pmatrix}, \tag{2.7}$$

$$\begin{pmatrix} R_t^{k,y} \\ R_t^{k,CC} \\ \boldsymbol{f}_t^k \end{pmatrix} = \begin{pmatrix} R_{t-1}^{k,y} \\ R_{t-1}^{k,CC} \\ \boldsymbol{f}_{t-1}^k \end{pmatrix} + \begin{pmatrix} \eta_{R,t}^{k,y} \\ \eta_{R,t}^{k,CC} \\ \boldsymbol{u}_t^k \end{pmatrix}, \tag{2.8}$$

$$\operatorname{cov}\begin{pmatrix} \eta_{R,t}^{k,y} \\ \eta_{R,t}^{k,CC} \\ \boldsymbol{u}_t^k \end{pmatrix} = \begin{bmatrix} \sigma_{R,y}^2 & \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \rho_{1,GT}\sigma_{R,y} & \cdots & \rho_{r,GT}\sigma_{R,y} \\ \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \sigma_{R,CC}^2 & 0 & \ldots & 0 \\ \rho_{1,GT}\sigma_{R,y} & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{r,GT}\sigma_{R,y} & 0 & 0 & \ldots & 1 \end{bmatrix}. \tag{2.9}$$

The last equality allows the innovations of the trends' slopes, $R_t^{k,y}$ and $R_t^{k,CC}$, and of the factors of the Google Trends to be correlated. Harvey and Chung (2000) show that there can be potential gains in precision, in terms of $\operatorname{MSE}\left(\hat{\theta}_t^{k,y}\right)$, if the correlation parameters $|\rho|$s are large. Specifically, if $|\rho_{CC}| = 1$, then $\boldsymbol{y}_t^k$ and $x_t^{k,CC}$ have a common slope. This means that $\boldsymbol{y}_t^k$ and $x_t^{k,CC}$ are both $I(2)$, but there is a linear combination of their first differences which is stationary. Likewise, if $|\rho_{m,GT}| = 1$ then the m^{th} factor of the Google Trends and the change in unemployment, $R_t^{k,y}$, are cointegrated. This is why we need the elements of the vector in (2.8) to have the same order of integration, and it is via this correlation parameters that we exploit the auxiliary information.

We allow the factors of the Google Trends to be correlated with the change in unemployment and not with its level for two reasons: firstly, a smooth trend model is assumed for the population parameter, which means that the level of its trend does not have an innovation term. Secondly, it is reasonable to assume that people start looking for a job on the internet when they become unemployed, and hence their search behaviour should reflect the change in unemployment rather than its level.

The Kalman filter (second step) is applied to the whole state space model (equations (2.5)-(2.9)) to re-estimate $\boldsymbol{f}_t^k$ and to nowcast the variable of interest, $\theta_t^{k,y}$, providing unemployment estimates in real time before LFS data become available:

$$\hat{\theta}_t^{k,y} = \mathrm{E}\left[\theta_t^{k,y} | \Omega_t; \mathcal{M}_{\left(\hat{\Lambda},\hat{\Psi}\right)}\right]$$

is the now-cast for $\theta_t^{k,y}$.

Since in each week we can aggregate the weekly Google Trends to the monthly frequency, we can use the information available throughout the month to update the estimated factors and loadings of the auxiliary series. If the correlations between the factors and the trend's slope of the target variable are large, this update should provide a more precise nowcast of $R_t^{k,y}$, $L_t^{k,y}$ and $\theta_t^{k,y}$.

See Appendix A.1, A.2 and A.3 for a detailed state space representation of the labour force model when, respectively, a univariate, a high-dimensional or both type of auxiliary series are included.

Our method to include auxiliary information in a state space model is based on the approach proposed by Harvey and Chung (2000). The factors of the high-dimensional auxiliary series could also be included as regressors in the observation equation for the labour force. Nevertheless, in such a model, the main part of the trend $L_t^{k,y}$ will be explained by the auxiliary series in the regression component. As a result, the filtered estimates for $L_t^{k,y}$ will contain a residual trend instead of the trend of the unemployment. Since the filtered trend estimates are the most important target variables in the official monthly publications of the labour force, this approach is not further investigated in this paper.

2.2.3 Extensions

We consider three different extensions of the proposed high-dimensional state space model, in order to achieve better results for the nowcast of the unobserved components.

Targeting the Google Trends. Bai and Ng (2008) show that targeting the predictors with the Elastic Net before estimating their factors can improve their forecasting performance. We follow their approach and regress the differenced estimated change in unemployment from the labour force model without auxiliary series, $\Delta\hat{R}_t^{k,y}$, on the differenced Google Trends using the penalized regression proposed by Hastie and Zou (2005), which solves the following minimization problem:

$$\min_{\boldsymbol{\beta}} \left[\frac{1}{2T}\sum_{t=1}^{T}\left(\Delta\hat{R}_t^{k,y} - \boldsymbol{\beta}'\Delta\boldsymbol{x}_t^{k,GT}\right)^2 + \lambda P_\alpha\left(\boldsymbol{\beta}\right)\right],$$

where

$$P_\alpha\left(\boldsymbol{\beta}\right) = \left(1-\alpha\right)\frac{1}{2}||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1.$$

The tuning parameter λ is chosen in order to minimize the Akaike (1974) information criterion (AIC) for a grid of values of $\alpha \in [0.05, \ldots, 0.95]$.

Including the lags of the Google Trends factors. It is reasonable to assume that people might start looking for a job before becoming unemployed. We therefore propose a parsimonious method to let the innovation of the change in unemployment depend on the lags of the Google Trends factor. Assume only one relevant factor for the Google Trends and consider a regression of $\eta_{R,t}^{k,y}$ on the lags of the differenced factor:

$$\eta_{R,t}^{k,y} = \sum_{j=1}^{q}\kappa_j u_{t-j}^k + w_t^k = \kappa_1 f_{t-1}^k + \sum_{j=2}^{q}\left(\kappa_j - \kappa_{j-1}\right)f_{t-j}^k - \kappa_q f_{t-q-1}^k + w_t^k, \quad w_t^k \sim N\left(0, \sigma_w^2\right).$$

$\eta_{R,t}^{k,y}$ is estimated from the labour force model without auxiliary series, and regressed on $\hat{u}_t^k$, estimated by PCA, in order to obtain ordinary least squares estimates of the parameters κ. The choice of the number of lags to be included in the regression is chosen by the AIC. The estimated κ are then incorporated in the transition equation as described in Appendix A.2.1.

Modelling the seasonality/cycle of the Google Trends' factors. The Google Trends factors might capture cycles or the seasonality of the job search terms. Assume again only one relevant factor for the Google Trends. In line with Alonso et al. (2011), instead of deseasonalizing the Google Trends, we model the seasonality of the factor as follows. We assume that f_t^k follows a seasonal ARIMA model $(p,d,q) \times (P,D,Q)_s$:

$$(1-B)^d(1-B^s)^D\phi(B)\Phi(B^s)f_t^k = \gamma(B)\Gamma(B^s)u_t^k, \quad u_t^k \sim N(0,1),$$

where B is the lag operator. Once f_t^k is estimated by PCA, the parameters of the seasonal ARIMA model, ϕ and γ, can be estimated by ordinary least squares and plugged in the transition equation of the state space model, in a similar fashion as proposed for including the lags of the factor (see Appendix A.2.2 for details on the state space representation of an ARIMA$(3,1,1)$ specification for the factor). If $s=0$, then this method models the cycle, instead of the seasonality, of the factor. We do not model the seasonality of the factor with a trigonometric seasonal component, as done for the population parameter in equation (2.1), because it would not allow us to model either the seasonality or the cycle of the factor, but only the former.

3 Simulations

We next conduct a Monte Carlo simulations study in order to elucidate to which extent our proposed method can provide gains in the nowcast accuracy of the unobserved components of interest. For this purpose, we consider a simpler model than the one used for the labour force survey. Namely, y_t^k is univariate and follows a smooth trend model. $\boldsymbol{x}_t^k$ represents the (100×1)-dimensional auxiliary series and has one common factor $(r=1)$.

$$\begin{aligned}
\begin{pmatrix} y_t^k \\ \boldsymbol{x}_t^k \end{pmatrix} &= \begin{pmatrix} L_t^k \\ \Lambda f_t^k \end{pmatrix} + \begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix}, \\
L_t^k &= L_{t-1}^k + R_{t-1}^k, \\
\begin{pmatrix} R_t^k \\ f_t^k \end{pmatrix} &= \begin{pmatrix} R_{t-1}^k \\ f_{t-1}^k \end{pmatrix} + \begin{pmatrix} \eta_{R,t}^k \\ u_t^k \end{pmatrix}, \quad \begin{pmatrix} \eta_{R,t}^k \\ u_t^k \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).
\end{aligned}$$

We allow the slope and factor's innovations to be correlated, and we investigate the performance of the model for increasing values of the correlation parameter $\rho \in [0, 0.2, 0.4, 0.6, 0.8, 0.9, 0.99]$. $\boldsymbol{x}_t^k$ has the same frequency of y_t^k and it is assumed that all $\boldsymbol{x}_t^k$ are released at the same time without publication delays. The nowcast is done concurrently, i.e. in realtime. This means that in each time point of the out-of-sample period, the hyperparameters of the model are re-estimated by maximum likelihood. This is done in the third part of the sample, always assuming that y_t^k is not available at time t, contrary to $\boldsymbol{x}_t^k$. The sample size is $T=150$ and the number of simulations is $n_{\text{sim}}=500$.

We consider three specifications for the idiosyncratic components and the factor loadings:

1. Homoscedastic idiosyncratic components and dense loadings:

$$\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N(\mathbf{0}, \operatorname{diag}(0.5)), \quad \Lambda \sim U(0,1).$$

2. Homoscedastic idiosyncratic components and sparse loadings. The first half of the elements in the loadings are set equal to zero. This specification reflects the likely empirical case that some of the Google Trends are not related to the change in unemployment:

$$\begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N\left(\mathbf{0}, \mathrm{diag}(0.5)\right), \quad \Lambda = \left(\Lambda_0', \Lambda_1'\right)', \underset{50\times 1}{\Lambda_0} = \mathbf{0}, \underset{50\times 1}{\Lambda_1} \sim U\left(0,1\right).$$

3. Heteroscedastic idiosyncratic components and dense loadings. The homoscedasticity assumption is here relaxed, again as not being realistic for the job search terms:

$$H \sim U(0.5, 10), \begin{pmatrix} \varepsilon_t^{k,y} \\ \boldsymbol{\varepsilon}_t^{k,x} \end{pmatrix} \sim N\left(\mathbf{0}, \mathrm{diag}(H)\right), \quad \Lambda \sim U\left(0,1\right).$$

Let $\boldsymbol{\alpha}_t^k = \left(L_t^k, R_t^k, f_t^k\right)'$ denote the vector of state variables. The results from the Monte Carlo simulations are shown in Table 2. We always report the MSFE, together with its variance and bias components, of the Kalman filter estimator of $\boldsymbol{\alpha}_t^k$, relative to the same measures calculated from the model that does not include the auxiliary series $\boldsymbol{x}_t^k$.

$$\begin{aligned}
\mathrm{MSFE}(\hat{\boldsymbol{\alpha}}_t^k) &= \frac{1}{h}\sum_{t=T-h+1}^{T} \frac{1}{n_{\mathrm{sim}}}\sum_{j=1}^{n_{\mathrm{sim}}} \left(\hat{\boldsymbol{\alpha}}_{jt} - \boldsymbol{\alpha}_{jt}\right)\left(\hat{\boldsymbol{\alpha}}_{jt} - \boldsymbol{\alpha}_{jt}\right)', \\
\mathrm{var}(\hat{\boldsymbol{\alpha}}_t^k) &= \frac{1}{h}\sum_{t=T-h+1}^{T} \left(\frac{1}{n_{\mathrm{sim}}}\sum_{j=1}^{n_{\mathrm{sim}}} \left(\left(\hat{\boldsymbol{\alpha}}_{jt} - \boldsymbol{\alpha}_{jt}\right) - \frac{1}{n_{\mathrm{sim}}}\sum_{j=1}^{n_{\mathrm{sim}}} \left(\hat{\boldsymbol{\alpha}}_{jt} - \boldsymbol{\alpha}_{jt}\right) \right) \right. \\
&\qquad \left. \times \left(\left(\hat{\boldsymbol{\alpha}}_{jt} - \boldsymbol{\alpha}_{jt}\right) - \frac{1}{n_{\mathrm{sim}}}\sum_{j=1}^{n_{\mathrm{sim}}} \left(\hat{\boldsymbol{\alpha}}_{jt} - \boldsymbol{\alpha}_{jt}\right) \right)' \right), \\
\mathrm{bias}^2(\hat{\boldsymbol{\alpha}}_t^k) &= \frac{1}{h}\sum_{t=T-h+1}^{T} \left(\frac{1}{n_{\mathrm{sim}}}\sum_{j=1}^{n_{\mathrm{sim}}} \left(\hat{\boldsymbol{\alpha}}_{jt} - \boldsymbol{\alpha}_{jt}\right) \right) \left(\frac{1}{n_{\mathrm{sim}}}\sum_{j=1}^{n_{\mathrm{sim}}} \left(\hat{\boldsymbol{\alpha}}_{jt} - \boldsymbol{\alpha}_{jt}\right) \right)',
\end{aligned}$$

where h is the size of the of out-of-sample period.

In every setting, both the bias and the variance of the MSFE tend to decrease with the magnitude of the correlation parameter. The improvement is more pronounced for the slope rather than the level of the trend. For the largest value of the correlation, with respect to the model which does not include auxiliary information, the gain in MSFE for the level and the slope is, respectively, of around 25% and 75%. Moreover, for low values of ρ, the MSFE does not deteriorate with respect to the benchmark model. This implies that our proposed method is robust to the inclusion of auxiliary information that does not have predictive power for the state variables of interest.

4 Empirics

As explained in Section 2.2, the Google series used in the model must be $I(1)$. We therefore test for nonstationarity in the Google Trends with the Elliott et al. (1996) augmented Dickey-Fuller (ADF) test, including a constant and a linear trend. We control for the false discovery rate as in Moon and Perron (2012), who employ a moving block bootstrap approach that accounts for time and cross-sectional dependence among the units in the panel. We proceed with the estimation of the model by only including the Google

	$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$	$\rho = 0.9$	$\rho = 0.99$
	Homoscedastic idiosyncratic components and dense loadings						
$\text{MSFE}(\hat{L}_t^k)$	1.030	1.024	1.006	0.971	0.901	0.837	0.718
$\text{var}(\hat{L}_t^k)$	1.031	1.025	1.007	0.971	0.901	0.837	0.718
$\text{bias}^2(\hat{L}_t^k)$	0.775	0.767	0.756	0.733	0.692	0.659	0.567
$\text{MSFE}(\hat{R}_t^k)$	1.044	1.017	0.941	0.806	0.588	0.427	0.198
$\text{var}(\hat{R}_t^k)$	1.045	1.018	0.942	0.807	0.589	0.427	0.198
$\text{bias}^2(\hat{R}_t^k)$	0.650	0.633	0.583	0.492	0.350	0.252	0.122
	Homoscedastic idiosyncratic components and sparse loadings						
$\text{MSFE}(\hat{L}_t^k)$	1.031	1.026	1.011	0.981	0.920	0.862	0.744
$\text{var}(\hat{L}_t^k)$	1.031	1.026	1.012	0.981	0.920	0.862	0.745
$\text{bias}^2(\hat{L}_t^k)$	0.784	0.776	0.762	0.737	0.695	0.655	0.582
$\text{MSFE}(\hat{R}_t^k)$	1.044	1.019	0.946	0.817	0.605	0.446	0.208
$\text{var}(\hat{R}_t^k)$	1.045	1.020	0.947	0.817	0.606	0.446	0.209
$\text{bias}^2(\hat{R}_t^k)$	0.656	0.639	0.586	0.492	0.347	0.243	0.104
	Heteroscedastic idiosyncratic components and dense loadings						
$\text{MSFE}(\hat{L}_t^k)$	1.036	1.032	1.019	0.994	0.945	0.901	0.823
$\text{var}(\hat{L}_t^k)$	1.037	1.032	1.020	0.995	0.946	0.902	0.823
$\text{bias}^2(\hat{L}_t^k)$	0.707	0.645	0.579	0.521	0.484	0.483	0.543
$\text{MSFE}(\hat{R}_t^k)$	1.049	1.027	0.960	0.840	0.644	0.499	0.299
$\text{var}(\hat{R}_t^k)$	1.049	1.028	0.961	0.841	0.645	0.500	0.299
$\text{bias}^2(\hat{R}_t^k)$	0.805	0.697	0.556	0.397	0.230	0.161	0.237

Table 2: Simulation results from the three settings described in Section 3. The values are reported relative to the respective measures calculated from the model that does not include the auxiliary series; values < 1 are in favour of our method. $n_{\text{sim}} = 500$.

Trends that resulted as being $I(1)$ from the multiple hypotheses testing. The number of nonstationary Google Trends, n, may differ depending on whether the weekly Google Trends have been aggregated according to equation (2.3) or (2.4), as we test for nonstationary after the temporal aggregation, or if the monthly Google Trends are used. Whenever we apply PCA or the Elastic Net, the Google Trends are first differenced and standardized.

We need to make sure that the stationarity assumption of the idiosyncratic components is maintained. Therefore, after having estimated the factors by PCA in (2.2), we test which of the idiosyncratic components $\boldsymbol{\varepsilon}_t$ are $I(1)$ with an ADF test without deterministic components, by controlling for multiple hypotheses testing as in Moon and Perron (2012). The $I(1)$ idiosyncratic components are modelled as state variables in (2.5), with the following transition equation:

$$\boldsymbol{\varepsilon}_t^k = \boldsymbol{\varepsilon}_{t-1}^k + \boldsymbol{\xi}_t^k,$$

with usual normality assumptions on the $\boldsymbol{\xi}_t^k$. The covariance matrix of the idiosyncratic components $\boldsymbol{\Psi}$ is therefore estimated on the levels of the $I(0)$ idiosyncratic components and the first differences of the $I(1)$ idiosyncratic components (which are around half of them). Appendix A.2.3 provides a toy example that elucidates the estimation procedure.

We always estimate four different models: the labour force model without auxiliary series (baseline), the labour force model with auxiliary series of claimant counts (CC), of Google Trends (GT) and of both (CC & GT). We compare the latter three models to the baseline one with an in-sample and an out-of-sample exercise. The period considered for the estimation starts in January 2004 and ends in December 2017 ($T = 167$ months). The out-of-sample nowcasts are conducted in real time (concurrently) in the third part of the sample; each week or month, depending on whether we use weekly or monthly Google Trends, the model is re-estimated assuming that the current observations for the unemployed labour force and the claimant counts are missing.

We define the measure of estimation accuracy $\widehat{\text{MSE}}(\hat{\boldsymbol{\alpha}}_t^k) = \frac{1}{T-d}\sum_{t=d+1}^{T} \hat{\boldsymbol{P}}_{t|t}^k$, where $\boldsymbol{\alpha}_t^k$ is the vector of state variables, $\hat{\boldsymbol{P}}_{t|t}^k$ is its estimated covariance matrix in month t, and d is the number of nonstationary state variables that are needed to estimated the labour force model without auxiliary series (i.e., 19). The measure of nowcast accuracy, $\widehat{\text{MSFE}}(\hat{\boldsymbol{\alpha}}_t^k) = \frac{1}{h}\sum_{t=T-h+1}^{T} \hat{\boldsymbol{P}}_{t|t}^k$, is the average of the nowcasted covariance matrices in the h prediction months. When weekly Google Trends are used, $\hat{\boldsymbol{P}}_{t|t}^k = \frac{1}{k}\sum_{j=1}^{k} \hat{\boldsymbol{P}}_{t|j}^k$, where $\hat{\boldsymbol{P}}_{t|j}^k$ is the nowcasted covariance matrix for the prediction in week j of month t. This is because the nowcast is done recursively throughout the weeks of the out-of-sample period. We always report the relative $\widehat{\text{MS(F)E}}$ with respect to the baseline model; values lower than one are in favour of our method.

The initial parameters for the maximum likelihood estimation are equal to the estimates for the labour force model in van den Brakel and Krieg (2015). The hyperparameter estimates for the survey errors are divided by (1 - $\hat{\delta}^2$), which implies that δ is treated as known and replaced by $\hat{\delta} = 0.21$ (again from van den Brakel and Krieg (2015)) in the estimation. We use a diffuse initialisation of the Kalman filter for all the nonstationary state variables (except for the 13 state variables that define the autocorrelation structure of the survey errors for which we use the exact initialisation of van den Brakel and Krieg (2016)).

Scree plots and the majority of the information criteria proposed by Bai and Ng (2002) agree in estimating only one factor from the Google Trends, no matter whether they are observed on a weekly or a monthly basis. This result is consistent with the fact that these job search terms should only explain the change in unemployment, and hence be driven by one common factor. Hence, the covariance matrix in (2.9) has only two correlation parameters: ρ_{CC} and $\rho_{1,GT}$, which we denote with ρ_{GT} in the remainder of the paper.

We conduct a Wilks (1938) likelihood ratio (LR) test to assess whether the correlation parameters are different from zero, and hence adding the auxiliary information might yield a significant improvement from the baseline model. Namely, the null hypotheses of the test for the CC, GT and CC & GT models are, respectively: $\rho_{CC} = 0$, $\rho_{GT} = 0$ and $\rho_{CC} = \rho_{GT} = 0$. The test statistics should be compared to the critical values of a χ^2 distribution with degrees of freedom equal to the number of parameters that are being tested.

Table 3 reports the estimated parameters for the four models as well as the relative measures of in and out-of-sample performance when the monthly Google Trends are used. We consider two different specifications of the GT model: the one where the innovation of the factor has variance fixed $\sigma_u = 1$, for identifiability reasons; the other one where σ_u is estimated by maximum likelihood, to give more flexibility to the model (see Appendix A.2 for the details of the model's specification in the latter case). As the results do not differ too much, we decide to keep $\sigma_u = 1$. The estimated correlation with the claimant counts is large, more than 0.9, and remains such when including the Google Trends. On the contrary, the correlation with the Google Trends' factor slightly decreases with respect to the GT model, where it is around 0.03.[1] Hence, the series of claimant counts has such a strong explanatory power for the unemployment, that it annihilates the contribution of the Google Trends. The best out-of-sample results, in terms of nowcast accuracy of all the state variables, are achieved for the models that contain the claimant counts, with a gain of around 7% for $\hat{R}_t^{k,y}$ and almost 30% for $\hat{L}_t^{k,y}$ and $\hat{\theta}_t^{k,y}$. When only the Google Trends are included, the

[1] The sign of the correlation with the factor is not relevant as the factor, in PCA, is identified up to a sign. We are only interested in the magnitude of this parameter.

measures of estimation and nowcast accuracy are of the same magnitude as those for the baseline model. The LR tests suggest that only the correlation between the unemployment's change and the slope of the claimant counts is significantly different from zero, indicating a preference for the model which contains this auxiliary information rather than the Google Trends.

			$\sigma_u = 1$		σ_u not fixed	
	Baseline	CC	GT	CC & GT	GT	CC & GT
$\hat{\sigma}_{R,y}$	2201.140	3023.983	2206.526	3011.524	2196.600	3029.661
$\hat{\sigma}_{\omega,y}$	0.020	0.020	0.020	0.020	0.020	0.020
$\hat{\sigma}_{\lambda}$	1166.055	1214.057	919.196	969.078	917.883	1095.453
$\hat{\sigma}_{\nu_1}$	1.165	1.169	1.171	1.170	1.171	1.170
$\hat{\sigma}_{\nu_2}$	1.139	1.139	1.143	1.141	1.139	1.139
$\hat{\sigma}_{\nu_3}$	1.082	1.077	1.081	1.082	1.087	1.078
$\hat{\sigma}_{\nu_4}$	1.128	1.144	1.129	1.146	1.129	1.145
$\hat{\sigma}_{\nu_5}$	1.100	1.107	1.098	1.109	1.098	1.106
$\hat{\sigma}_{R,CC}$		3606.933		3605.960		3602.868
$\hat{\sigma}_{\omega,CC}$		0.020		0.020		0.020
$\hat{\sigma}_{\varepsilon,CC}$		1120.032		1120.562		1128.259
$\hat{\sigma}_{u}$					1.032	1.031
$\hat{\rho}_{CC}$		0.902		0.903		0.905
$\hat{\rho}_{GT}$			0.025	0.013	0.038	-0.001
$\widehat{\text{MSE}}(\hat{L}_t^{k,y})$		0.869	0.992	0.857	0.991	0.860
$\widehat{\text{MSE}}(\hat{R}_t^{k,y})$		0.956	1.003	0.948	0.996	0.954
$\widehat{\text{MSE}}(\hat{\theta}_t^{k,y})$		0.890	0.993	0.881	0.993	0.883
$\widehat{\text{MSFE}}(\hat{L}_t^{k,y})$		0.715	0.997	0.707	0.996	0.704
$\widehat{\text{MSFE}}(\hat{R}_t^{k,y})$		0.929	1.003	0.932	1.004	0.929
$\widehat{\text{MSFE}}(\hat{\theta}_t^{k,y})$		0.729	0.997	0.721	0.996	0.718
		p-values from the LR test				
$H_0 : \rho_{CC} = 0$		0.0004		0.0004		0.0003
$H_0 : \rho_{GT} = 0$			1	1	1	1
$H_0 : \rho_{CC} = \rho_{GT} = 0$				0.0018		0.0000

Table 3: Estimation and nowcast results for the labour force model with and without auxiliary series. The auxiliary series are the claimant counts and the $n = 74$ monthly Google Trends about job search terms.

In Table 4 we report the estimation and nowcast results for the models which employ the weekly Google Trends. We consider both types of temporal aggregation of the Google Trends, given by equations (2.3) and (2.4). Moreover, in section 2.2.1 we mention that the two-step estimation is done on the weekly data, which means that the factor loadings, Λ, and the covariance matrix of the idiosyncratic components, $\boldsymbol{\Psi}$, are also estimated from the weekly Google Trends. We believe that this type of estimation is more accurate as the sample size is larger than when using the monthly data. Furthermore, the high frequency observations are likely more informative about the dynamics of employment decisions of people searching for a job. Nevertheless, we report the results also when implementing the two-step estimation with the low-frequency search terms.

When Λ and $\boldsymbol{\Psi}$ are estimated on the weekly data, the correlation with the Google Trends' factor in the GT model is above 0.25, hence larger with respect to the data used in Table 3. In this case we still notice the loss of explanatory power of the Google Trends when the claimant counts are also included in the model, in terms of magnitude of the respective estimated correlation. The opposite effect is instead registered when Λ and $\boldsymbol{\Psi}$ are estimated on the monthly data.

The conclusions from the LR tests are the same as those from Table 3. However, when estimating the factor loadings and the covariance matrix of the idiosyncratic components with the weekly Google Trends aggregated according to equation (2.4), we obtain the best nowcast accuracy for all the state variables. In general, the measures of in and out-of-sample performance of the GT model do not differ much from the baseline one, with relative figures remaining slightly but broadly below 1.

The precision of the nowcast does not monotonically improve with the number of weeks. If the high-dimensional state space model could be expressed and estimated on the highest frequency, the weekly gains in nowcast accuracy could be more evident. Nonetheless, we are limited by the transition equations for the RGB and the survey errors, to estimate the model on the monthly frequency.

Next, we consider the extensions to the GT and CC & GT models proposed in Section 2.2.3 under the setting of the last two columns of Table 4, as it yields the best results in terms of nowcast accuracy of the change in unemployment. The results from these extensions are reported in Table 5. Targeting the Google Trends with the Elastic Net does not increase the value of the correlation parameter with their factor, nor significantly improves the in and out-of-sample performance of the models. Figure 5 shows the frequency of selection of the Google Trends in the out-of-sample period. The most selected search terms are (translated from Dutch): "uwv (Employee Insurance Agency) ww (Employment Insurance Act)", "to write an application letter", "start people (recruitment agency)", "randstad (recruitment agency) vacancies", "to request benefit", "job search", "uwv benefit", "unemployment benefit", "to retrain", and "volunteer". The number of Google Trends included in the model, n, varies between 8 and 78.

When including the lags of the factor, only one lag is always chosen by the AIC in each recursion of the out-of-sample exercise, and its estimated parameter is insignificant [2]. This already suggests that the inclusion of the lag should not improve the accuracy of the estimation. Indeed, the in and out-of-sample accuracy results slightly worsen, with respect to the baseline model and the model that includes only the claimant counts.

The estimated factor shows a cyclical pattern, especially when the factor loadings are estimated on the weekly Google Trends. Since the seasonal ARIMA model is fitted on the estimated monthly factor by PCA, $s = 12$. The number of lags and MA components are again chosen according to the AIC, which suggests an ARIMA$(3, 1, 1)$ as best model:

$$\Delta f_t^k = \phi_1 \Delta f_{t-1}^k + \phi_2 \Delta f_{t-2}^k + \phi_3 \Delta f_{t-3}^k + u_t^k + \gamma u_{t-1}^k,$$

suggesting quarterly dependence in the factor. Table 5 shows that for the CC & GT model, this extension yields a slight improvement in the in and out-of-sample estimation of $L_t^{k,y}$ and $\theta_t^{k,y}$, with respect to the CC model. The LR test is still not in favour of the model which also includes the Google Trends, but the inclusion of the search terms as only auxiliary information again does not worsen the measures of accuracy with respect to the baseline model.

Finally, since one out of the three panel criteria proposed in Bai and Ng (2002) suggests that the number of relevant factors is equal to two, we estimate the state space model with two factors for the Google Trends. We are here only interested in the in and out-of-sample performance of the model rather than finding the correct specification for the factors, so we do not employ an ARIMA specification for the factors. Table 5 shows that this is the only setting under which the GT model clearly outperforms the baseline, with

[2] The AIC would actually prefer to not include any lag, but we force at least one lag to be included in order to have a different model.

	aggregation (2.3), $n = 82$				aggregation (2.4), $n = 79$			
	monthly $\hat{\Lambda}$, $\hat{\boldsymbol{\Psi}}$		weekly $\hat{\Lambda}$, $\hat{\boldsymbol{\Psi}}$		monthly $\hat{\Lambda}$, $\hat{\boldsymbol{\Psi}}$		weekly $\hat{\Lambda}$, $\hat{\boldsymbol{\Psi}}$	
	GT	CC & GT	GT	CC & GT	GT	CC & GT	GT	CC & GT
$\hat{\sigma}_{R,y}$	2251.799	2671.998	2378.318	3049.543	2236.951	2966.467	2245.709	3034.202
$\hat{\sigma}_{\omega,y}$	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020
$\hat{\sigma}_{\lambda}$	929.098	1047.419	1168.364	1216.272	925.020	983.072	1168.075	1203.600
$\hat{\sigma}_{\nu_1}$	1.171	1.152	1.164	1.170	1.164	1.169	1.163	1.169
$\hat{\sigma}_{\nu_2}$	1.141	1.154	1.138	1.139	1.145	1.154	1.139	1.139
$\hat{\sigma}_{\nu_3}$	1.093	1.076	1.082	1.076	1.086	1.078	1.082	1.076
$\hat{\sigma}_{\nu_4}$	1.125	1.149	1.130	1.144	1.134	1.147	1.130	1.144
$\hat{\sigma}_{\nu_5}$	1.104	1.114	1.099	1.108	1.101	1.116	1.100	1.107
$\hat{\sigma}_{R,CC}$		3537.654		3613.233		3534.111		3609.248
$\hat{\sigma}_{\omega,CC}$		0.020		0.020		0.020		0.020
$\hat{\sigma}_{\varepsilon,CC}$		1132.106		1114.930		1176.415		1118.105
$\hat{\rho}_{CC}$		0.854		0.906		0.872		0.902
$\hat{\rho}_{GT}$	0.035	0.155	0.397	-0.045	-0.049	-0.144	0.254	-0.015
$\widehat{\mathrm{MSE}}(\hat{L}_t^{k,y})$	1.003	0.852	0.997	0.865	0.996	0.878	0.995	0.869
$\widehat{\mathrm{MSE}}(\hat{R}_t^{k,y})$	1.036	0.818	0.993	0.958	1.024	0.956	0.982	0.960
$\widehat{\mathrm{MSE}}(\hat{\theta}_t^{k,y})$	1.002	0.881	0.998	0.887	0.997	0.900	0.996	0.890
$\widehat{\mathrm{MSFE}}(\hat{L}_t^{k,y})$	0.979	0.823	0.991	0.714	0.986	0.744	0.988	0.709
week 1	0.984	0.816	0.994	0.717	0.992	0.728	0.989	0.707
week 2	0.976	0.832	0.990	0.719	0.981	0.751	0.987	0.712
week 3	0.973	0.830	0.991	0.708	0.985	0.754	0.989	0.709
week 4	0.985	0.815	0.993	0.718	0.987	0.753	0.989	0.713
week 5	0.970	0.820	0.981	0.693	0.979	0.718	0.977	0.691
$\widehat{\mathrm{MSFE}}(\hat{R}_t^{k,y})$	0.982	0.942	0.983	0.933	0.989	0.923	0.974	0.925
week 1	0.998	0.932	0.989	0.932	1.000	0.925	0.975	0.926
week 2	0.975	0.953	0.978	0.935	0.977	0.928	0.972	0.925
week 3	0.964	0.931	0.982	0.933	0.986	0.919	0.974	0.927
week 4	0.992	0.952	0.986	0.933	0.990	0.922	0.976	0.923
week 5	0.982	0.946	0.976	0.929	0.996	0.920	0.969	0.924
$\widehat{\mathrm{MSFE}}(\hat{\theta}_t^{k,y})$	0.981	0.832	0.992	0.727	0.987	0.756	0.989	0.723
week 1	0.985	0.825	0.994	0.730	0.993	0.741	0.990	0.721
week 2	0.978	0.841	0.991	0.732	0.982	0.763	0.988	0.725
week 3	0.975	0.838	0.992	0.722	0.986	0.766	0.990	0.723
week 4	0.987	0.823	0.994	0.731	0.988	0.765	0.990	0.727
week 5	0.972	0.829	0.982	0.707	0.980	0.731	0.978	0.705
	p-values from the LR test							
$H_0 : \rho_{CC} = 0$		0.0004		0.0004		0.0004		0.0004
$H_0 : \rho_{GT} = 0$	1	1	0.5271	1	1	1	0.5485	1
$H_0 : \rho_{CC} = \rho_{GT} = 0$		0.0021		0.0017		0.0019		0.0017

Table 4: Estimation and nowcast results for the labour force model with auxiliary series of claimant counts and aggregated weekly Google Trends to the monthly frequency. $\sigma_u = 1$ is fixed. The first type of aggregation is done with equation (2.3), whereas the second one with equation (2.4). "Weekly $\hat{\Lambda}$, $\hat{\boldsymbol{\Psi}}$" means that the factor loadings and the covariance matrix of the idiosyncratic components are estimated on the weekly Google Trends rather than the aggregated monthly ones.

an improvement in the estimation and nowcast accuracy for the change in unemployment of almost 12%, compared to the 2.5% of including only one factor. Moreover, the correlation parameters with both factors

are close to 0.4 in absolute value. A slight improvement is also registered in the in and out-of-sample accuracy of the other two state variables. Nonetheless, the LR test is still not in favour of the inclusion of the Google Trends as auxiliary series, and the nowcast performance of the CC & GT model does not improve with respect to including only one factor. Figures 2-4 display the nowcast, respectively, of the change in unemployment, its level and the population parameter, when two factors of the Google Trends are included. Especially from the first graph, it is evident that the models including the claimant counts slightly deviate from the baseline model, which, on the contrary, gives similar results as those of the GT model. The reason is likely due to the large correlation coefficient with the claimant counts series, whose trend's slope drives $R_t^{k,y}$. This is not the case for the GT model as the correlations with the search terms' factors are not significantly different from zero.

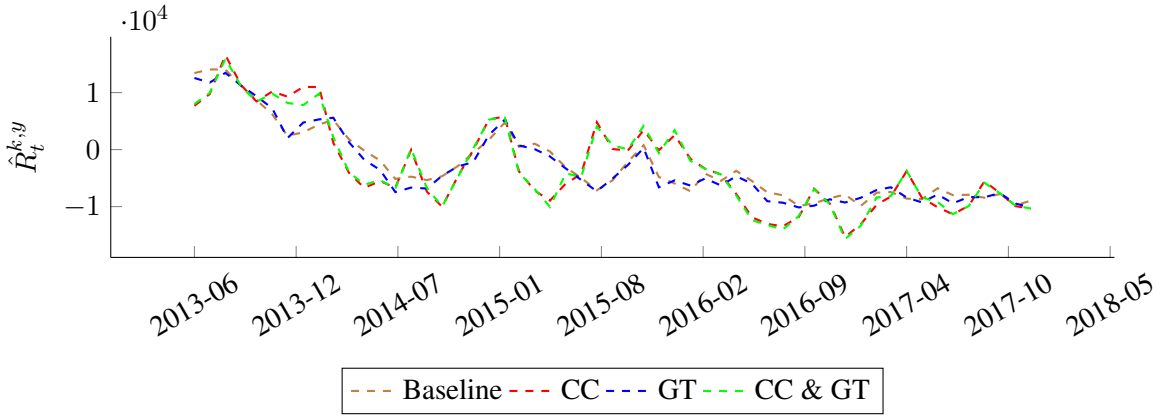

Figure 2: Nowcast of $R_t^{k,y}$ with the labour force models. The results for the GT and the CC & GT models refer to the setting where the seasonality of the factor is modelled. Each monthly value corresponds to the nowcast obtained in the last week of the month.

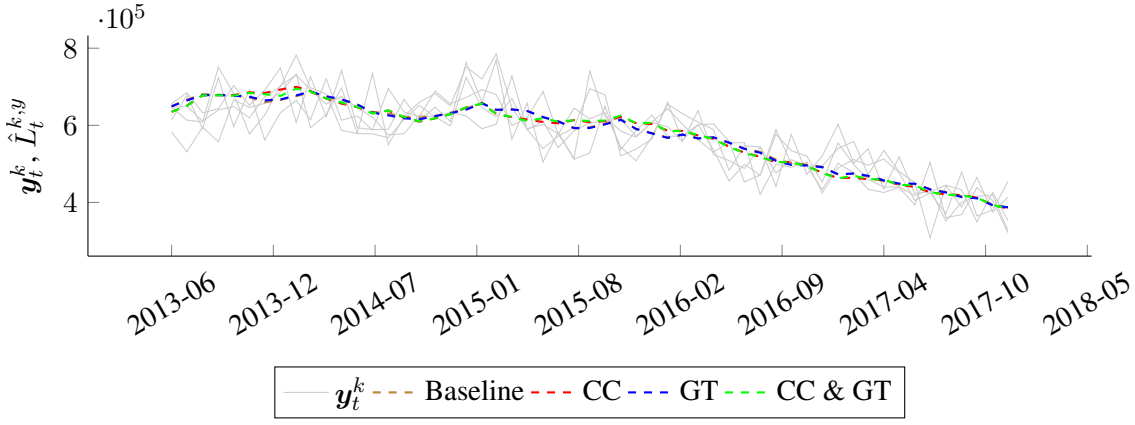

Figure 3: Nowcast of $L_t^{k,y}$ with the labour force models, compared to the five waves of the unemployed labour force. The results for the GT and the CC & GT models refer to the setting where the seasonality of the factor is modelled. Each monthly value corresponds to the nowcast obtained in the last week of the month.

	Elastic Net		1 lag of the factor		cycle of the factor		2 factors	
	GT	CC & GT	GT	CC & GT	GT	CC & GT	GT	CC & GT
$\hat{\sigma}_{R,y}$	2201.895	3044.793	2386.119	3035.427	2226.465	3084.042	2473.391	3096.768
$\hat{\sigma}_{\omega,y}$	0.020	0.020	0.020	0.020	0.020	0.020	0.020	0.020
$\hat{\sigma}_{\lambda}$	926.003	1052.668	1048.771	1205.303	935.002	1009.678	1037.313	1114.421
$\hat{\sigma}_{\nu_1}$	1.166	1.170	1.176	1.177	1.164	1.172	1.164	1.169
$\hat{\sigma}_{\nu_2}$	1.143	1.146	1.138	1.137	1.148	1.144	1.143	1.141
$\hat{\sigma}_{\nu_3}$	1.085	1.076	1.084	1.076	1.084	1.078	1.087	1.078
$\hat{\sigma}_{\nu_4}$	1.132	1.143	1.127	1.137	1.133	1.145	1.135	1.144
$\hat{\sigma}_{\nu_5}$	1.099	1.109	1.093	1.101	1.102	1.110	1.102	1.108
$\hat{\sigma}_{R,CC}$		3601.678		3612.823		3620.507		3595.186
$\hat{\sigma}_{\omega,CC}$		0.020		0.020		0.020		0.020
$\hat{\sigma}_{\varepsilon,CC}$		1138.115		1116.664		1111.063		1.052
$\hat{\rho}_{CC}$		0.902		0.882		0.909		0.903
$\hat{\rho}_{1,GT}$	-0.007	0.014	0.280	-0.009	0.016	-0.041	0.428	-0.038
$\hat{\rho}_{2,GT}$							-0.397	0.051
$\widehat{\mathrm{MSE}}(\hat{L}_t^{k,y})$	0.991	0.864	1.017	0.895	0.996	0.858	0.967	0.869
$\widehat{\mathrm{MSE}}(\hat{R}_t^{k,y})$	1.001	0.967	1.068	1.003	1.018	0.970	0.893	0.988
$\widehat{\mathrm{MSE}}(\hat{\theta}_t^{k,y})$	0.993	0.886	1.015	0.912	0.996	0.881	0.977	0.889
$\widehat{\mathrm{MSFE}}(\hat{L}_t^{k,y})$	0.986	0.688	1.028	0.771	0.995	0.694	0.949	0.731
week 1	0.986	0.688	1.029	0.767	0.997	0.706	0.949	0.737
week 2	0.988	0.689	1.028	0.772	0.995	0.690	0.949	0.722
week 3	0.988	0.688	1.028	0.774	0.996	0.690	0.951	0.731
week 4	0.988	0.688	1.029	0.780	0.994	0.693	0.952	0.743
week 5	0.976	0.691	1.017	0.748	0.985	0.686	0.938	0.703
$\widehat{\mathrm{MSFE}}(\hat{R}_t^{k,y})$	0.966	0.959	1.089	0.972	0.996	0.930	0.882	0.930
week 1	0.964	0.962	1.091	0.966	1.000	0.927	0.881	0.924
week 2	0.967	0.962	1.089	0.969	0.996	0.932	0.881	0.941
week 3	0.967	0.955	1.088	0.973	0.998	0.930	0.883	0.936
week 4	0.968	0.958	1.091	0.981	0.991	0.932	0.887	0.922
week 5	0.961	0.956	1.082	0.967	0.993	0.931	0.873	0.927
$\widehat{\mathrm{MSFE}}(\hat{\theta}_t^{k,y})$	0.987	0.703	1.025	0.782	0.995	0.708	0.953	0.743
week 1	0.987	0.702	1.026	0.778	0.997	0.720	0.953	0.749
week 2	0.989	0.703	1.026	0.782	0.996	0.705	0.953	0.735
week 3	0.989	0.702	1.026	0.785	0.997	0.704	0.955	0.744
week 4	0.989	0.703	1.027	0.790	0.995	0.708	0.956	0.756
week 5	0.977	0.705	1.015	0.759	0.985	0.701	0.943	0.717
	p-values from the LR test							
$H_0 : \rho_{CC} = 0$		0.0004		0.001		0.0004		0.0007
$H_0 : \rho_{1,GT} = 0$	1	1	0.5376	1	1	0.8415	0.3897	1
$H_0 : \rho_{2,GT} = 0$							0.3125	1
$H_0 : \rho_{1,GT} = \rho_{2,GT} = 0$							0.5117	
$H_0 : \rho_{CC} = \rho_{1,GT} = 0$		0.0017		0.0039		0.0019		
$H_0 : \rho_{CC} = \rho_{1,GT} = \rho_{2,GT} = 0$								0.0052

Table 5: Estimation and nowcast results for the labour force model with auxiliary series of claimant counts and aggregated weekly Google Trends to the monthly frequency. The aggregation is done according to equation (2.4). $\sigma_u = 1$ is fixed. The factor loadings and the covariance matrix of the idiosyncratic components are estimated on the weekly Google Trends. "Elastic Net" means that the Google Trends included in the model are first selected with the Elastic Net. The lag of the Google Trends' factor and its cycle are modelled, respectively, as described in Appendix A.2.1 and A.2.2. When including the lag of the factor, we estimate σ_w, defined in Appendix A.2.1, instead of $\sigma_{R,y}$.

The assumptions of no serial correlation, heteroscedasticity and normality made throughout the paper can be tested on the standardized one-step ahead forecast error of each series: $\tilde{v}_t^k = v_t^k / \sqrt{F_t^k}$, $t = d +$

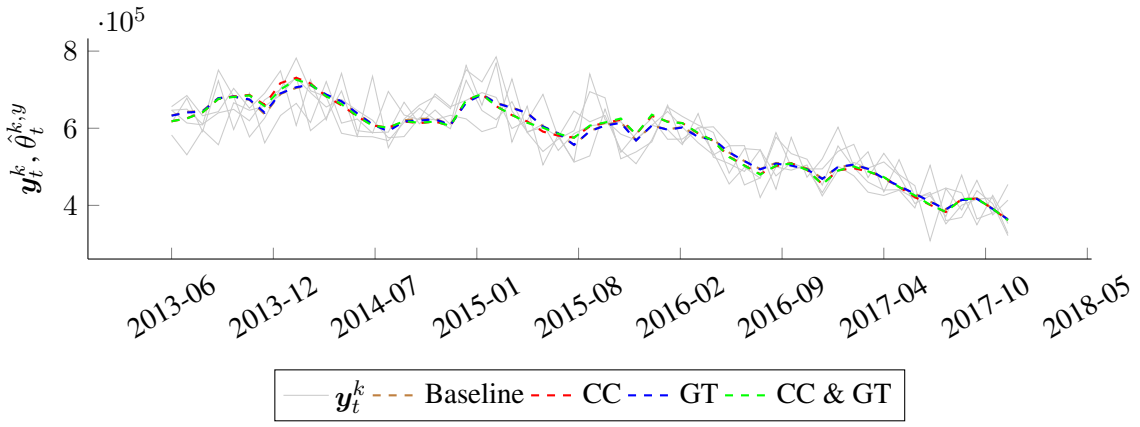

Figure 4: Nowcast of $\theta_t^{k,y}$ with the labour force models, compared to the five waves of the unemployed labour force. The results for the GT and the CC & GT models refer to the setting where the seasonality of the factor is modelled. Each monthly value corresponds to the nowcast obtained in the last week of the month.

$1, \ldots, T$. F_t^k is the variance of the prediction error v_t^k estimated with the Kalman filter. The forecast errors for the labour force are defined as $\boldsymbol{v}_t^{k,y} = \boldsymbol{y}_t^k - \boldsymbol{Z}^y \hat{\boldsymbol{\alpha}}_{t|t-1}^{k,y}$ and for the claimant counts as $v_t^{k,CC} = x_t^{k,CC} - \boldsymbol{Z}^{CC} \hat{\boldsymbol{\alpha}}_{t|t-1}^{k,CC}$. We test the assumptions on the estimated model with all auxiliary series, when two factors of the Google Trends are included. We test for serial correlation using the Ljung and Box (1978) test on 4, 8, 12 and 16 lags,[3] and for heteroscedasticity with the H(h) test, as suggested in Durbin and Koopman (2012). We test for univariate normality with the Shapiro and Wilk (1965) test as proposed by Harvey (1990).

The results from the diagnostic tests are shown in Table 6. When considering 4 and 12 lags, the Ljung-Box test rejects the null hypothesis of no serial correlation only for the claimant counts, at the 1% significance level. With 8 and 16 lags, the test again finds serially correlated forecast errors only for the claimant counts series at the 5% significance level due to seasonal patterns that are not captured by its seasonal component.[4] For the heteroscedasticity test we choose $h = 50$ and $h = 75$, which corresponds, respectively, to a third and a half of the sample size, $T - d$. The null hypothesis of homoscedasticity is never rejected at the 10% significance level for the labour force series, contrary to the claimant counts. The hypothesis of univariate normality is not rejected for any of the series at any critical level. We conclude that the sophisticated model for the labour force series is well specified, contrary to the model for the claimant counts. The fact that the auxiliary series we are using are not survey data, raises some challenges in understanding in which direction to improve their models. However, our main interest lies in the estimation of the population parameter and its components, which are part of the model for the unemployed labour force, and therefore in the correct specification of the latter. Table 7 shows that the results for the diagnostic tests are similar when only the claimant counts are included as auxiliary series.

If the idiosyncratic components have cross-sectional dependence, then the state space model is misspecified since we restrict $\hat{\boldsymbol{\Psi}}$ to be diagonal. Nonetheless, as long as this covariance matrix is invertible, only the efficiency, but not the consistency, of the Kalman filter estimator is affected. The same argument applies if the idiosyncratic components are autocorrelated, as we lose optimality but not consistency (see Barigozzi and Luciani (2017) for more details). Furthermore, even if the disturbances are non-normal, the forecasts of the state variables with the lowest mean squared errors are still provided by the Kalman filter (Hamilton, 1994). As misspecifications of the Google Trends are not of our concern, we do not carry out

[3] For each series, we correct the degrees of freedom of the test statistics, based on the number of parameters needed to estimate its unobserved components. See Harvey (1990) for more details on the degrees of freedom correction for the Ljung-Box test.

[4] The largest autocorrelations are registered every 6 lags.

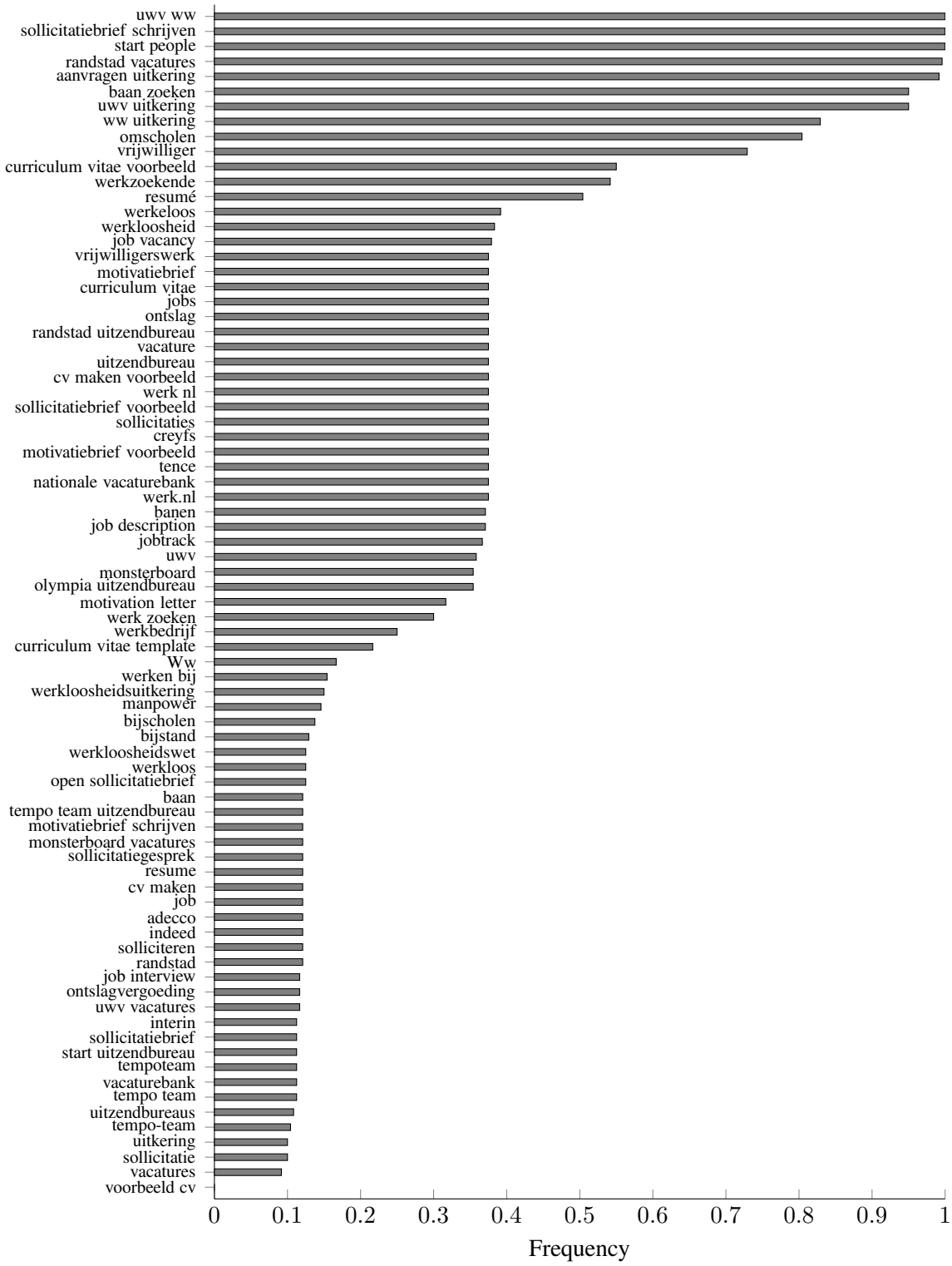

Figure 5: Frequency of Google search terms selection by the Elastic Net in the out-of-sample period. A value of 1 means that the variable has been selected in every week of the out-of-sample period.

the diagnostic tests on their prediction errors.

	LB(4)	LB(8)	LB(12)	LB(16)	H(50)	H(75)	SW
$\tilde{v}_t^{k,y_1}$	0.1692	0.6052	8.7728	0.5195	0.2729	0.3052	0.8319
$\tilde{v}_t^{k,y_2}$	0.1315	0.2516	0.1206	0.1653	0.4474	0.2747	0.1771
$\tilde{v}_t^{k,y_3}$	0.0228	0.2803	0.3404	0.4319	0.7327	0.5448	0.0830
$\tilde{v}_t^{k,y_4}$	0.0883	0.0955	0.0427	0.0677	0.9341	0.9804	0.3456
$\tilde{v}_t^{k,y_5}$	0.0492	0.4888	0.1265	0.3072	0.5841	0.6209	0.4729
$\tilde{v}_t^{k,CC}$	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.1421

Table 6: P-values of the diagnostic tests. LB(p) is the Ljung-Box test for serial correlation with $p = 4, 8, 12, 16$ lags. H(h) is the heteroscedasticity test. SW is the Shapiro-Wilk test for univariate normality. We test the assumptions on the estimated model including all auxiliary series, when two factors of the Google Trends are included.

	LB(4)	LB(8)	LB(12)	LB(16)	H(50)	H(75)	SW
$\tilde{v}_t^{k,y_1}$	0.1685	0.6037	0.8752	0.5091	0.2694	0.3000	0.8324
$\tilde{v}_t^{k,y_2}$	0.1344	0.2562	0.1242	0.1707	0.4531	0.2810	0.1841
$\tilde{v}_t^{k,y_3}$	0.0224	0.2782	0.3280	0.4167	0.7286	0.5350	0.0813
$\tilde{v}_t^{k,y_4}$	0.0875	0.0925	0.0442	0.0702	0.9336	0.9806	0.3288
$\tilde{v}_t^{k,y_5}$	0.0483	0.4803	0.1211	0.2997	0.5843	0.6217	0.4749
$\tilde{v}_t^{k,CC}$	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.1387

Table 7: P-values of the diagnostic tests. LB(p) is the Ljung-Box test for serial correlation with $p = 4, 8, 12, 16$ lags. H(h) is the heteroscedasticity test. SW is the Shapiro-Wilk test for univariate normality. We test the assumptions on the estimated model including only the claimant counts as auxiliary series.

5 Conclusions

This paper proposes a method to include a high-dimensional auxiliary series in a state space model in order to improve the estimation and nowcast of unobserved components. The method is based on a combination of PCA and Kalman filter estimations to reduce the dimensionality of the auxiliary series, originally proposed by Doz et al. (2011), while the auxiliary information is included in the state space model as in Harvey and Chung (2000).

In this way we extend the state space model used in Statistics Netherlands to estimate the Dutch unemployment, which is based on monthly LFS data, by including the auxiliary series of claimant counts and job-related Google searches. The strong explanatory power of the former series, in similar settings, has already been discovered in the literature (see Harvey and Chung (2000) and van den Brakel and Krieg (2016)). We explore to which extent a similar success can be obtained from online job-search behaviour. The advantage of Google Trends is that they are available at higher frequencies than the labour force survey and the claimant counts, and, contrary to the latter, they are not affected by publications delays. This feature can play a key role in the nowcast of the unemployment, as being the only real-time available information.

Results from a likelihood ratio test are in favour of a model that contains claimant counts rather than Google Trends. Nonetheless, the accuracy of the estimation and of the nowcast of the level and the change in unemployment, does not deteriorate when only the latter series are included. The measures of in and out-of-sample performance, with respect to the model which does not use auxiliary information, are indeed broadly lower than 1. We do not find great advantages in using weekly Google Trends over monthly ones. The gains are mainly due to more accurate estimations of the factor loadings and the covariance matrix of the idiosyncratic components. However, the monthly specification of the labour force survey model probably

prevents the full exploitation of the information coming from the higher frequency data. We estimate only one relevant factor from the Google Trends, and find that it captures the cyclical pattern of the search terms. The out-of-sample results slightly improve when the cycle of the factor is appropriately modelled according to an ARIMA model, and all the auxiliary series are included. Despite the intuition that job-related searches are performed before becoming unemployed, the change in unemployment does not seem to depend on the lagged Google Trends, suggesting that there is no explicit pattern in pre-emptive job-related searches that can be linked to a clear time span before becoming unemployed. Targeting the search terms with the Elastic Net before estimating the factors does not improve the estimation and nowcast accuracy of the unobserved components. Finally, including one additional factor of the search terms in the model improves its out-of-sample performance when the claimant counts are not included.

On the other hand, our Monte Carlo simulation study shows that in a smooth trend model our proposed method can improve the MSFE of the level and the slope of the trend up to, respectively, around 25% and 75%. Therefore, given the right auxiliary dataset, our method does have the potential to improve estimation and nowcasting of unobserved components, and it may simply be that the Google Trends series are not informative enough about unemployment.

Our choice of search terms is hand-picked, and therefore to some extent arbitrary and limited. We can therefore not rule out the usefulness of Google Trends for unemployment estimation in the Netherlands, although our results suggest limited scope for improvement. Clearly, for other topics or other countries results might be entirely different. Moreover, high-quality administration data such as the claimant counts are not available everywhere and for every series of interest. Finally, there are many other "Big Data" sources that could be considered for inclusion. Given that these will likely share the features of Google Trends of high-dimensionality, yet a low signal-to-noise ratio, such data sources can be treated similarly. Hence, our proposed approach provides a framework to analyse the usefulness of such sources as well, with little risk in case the series do not appear to be useful.

One limitation of the current paper is that it does not allow for time-variation in the relation between the unobserved component of interest and the auxiliary series. For example, legislative changes may change the correlation between unemployment and administrative series such as claimant counts. Additionally, one can easily imagine the relevance of both specific search terms as well as internet search behaviour overall to change over time. While such time-variation may partly be addressed by considering shorter time periods, decreasing the already limited time dimension will have a strong detrimental effect on the quality of the estimators. Therefore, a more structural method is required that extends the current approach by building the potential for time variation into the estimation method directly, while retaining the possibility to use the full sample size. Such extensions are currently under investigation by the authors.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alonso, A. M., Rodríguez, J., García-Martos, C., and Jesús Sánchez, M. (2011). Seasonal Dynamic Factor Analysis and Bootstrap Inference: Application to Electricity Market Forecasting. *Technometrics*, 53(2):137–151.

Askitas, N. and Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2):107–120.

Bai, J. (2004). Estimating Cross-section Common Stochastic Trends in Nonstationary Panel Data. *Journal of Econometrics*, 122(1):137–183.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2008). Forecasting Economic Time Series Using Targeted Predictors. *Journal of Econometrics*, 146(2):304–317.

Bailar, B. (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys. *Journal of the American Statistical Association*, 70(349):23–30.

Bańbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Now-casting and the Real-time Data Flow. Working Paper Series 1564, European Central Bank.

Barigozzi, M. and Luciani, M. (2017). Common Factors, Trends, and Cycles in Large Datasets. Finance and economics discussion series 2017-111, Board of Governors of the Federal Reserve System (U.S.).

Bollineni-Balabay, O., van den Brakel, J., and Palm, F. (2017). State Space Time Series Modelling of the Dutch Labour Force Survey: Model Selection and Mean Squared Errors Estimation. *Survey Methodology*, 43(1):41–67.

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*, 88(SUPPL.1):2–9.

Choi, H. and Varian, H. R. (2009). Predicting Initial Claims for Unemployment Benefits. *Google Inc*, pages 1–5.

Doz, C., Giannone, D., and Reichlin, L. (2011). A Two-step Estimator for Large Approximate Dynamic Factor Models Based on Kalman filtering. *Journal of Econometrics*, 164(1):188–205.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series. OUP Oxford.

Elliott, G., Rothenberg, T. J., and Stock, J. H. (1996). Efficient Tests for an Autoregressive Unit Root. *Econometrica*, 64(4):813–836.

Giannone, D., Reichlin, L., and Small, D. (2008). Nowcasting: The Real-time Informational Content of Macroeconomic Data. *Journal of Monetary Economics*, 55(4):665–676.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Harvey, A. and Chung, C.-H. (2000). Estimating the Underlying Change in Unemployment in the UK. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(3):303–309.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press.

Hastie, T. and Zou, H. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Hindrayanto, I., Koopman, S. J., and de Winter, J. (2016). Forecasting and Nowcasting Economic Growth in the Euro Area Using Factor Models. *International Journal of Forecasting*, 32(4):1284–1305.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). Supplementary Materials for The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(March):1203–1206.

Ljung, G. M. and Box, G. E. P. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 65(2):297–303.

Moon, H. R. and Perron, B. (2012). Beyond Panel Unit Root Tests: Using Multiple Testing to Determine the Nonstationarity Properties of Individual Series in a Panel. *Journal of Econometrics*, 169(1):29–33.

Pfeffermann, D. (1991). Estimation and Seasonal Adjustment of Population Means Using Data from Repeated Surveys. *Journal of Business and Economic Statistics*, 9(2):163–175.

Pfeffermann, D., Feder, M., and David, S. (1998). Estimation of Autocorrelations of Survey Errors with Application to Trend Estimation in Small Areas. *Journal of Business & Economic Statistics*, 16(3):339–348.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*. Wiley Series in Survey Methodology. John Wiley & Sons, Inc., 2 edition.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag Publishing, New York, NY, US.

Shapiro, S. S. and Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52(3-4):591–611.

Stephens-Davidowitz, S. and Varian, H. (2015). A Hands-on Guide to Google Data. *Google, Inc.*, pages 1–25.

Suhoy, T. (2009). Query Indices and a 2008 Downturn: Israeli Data. Discussion paper series no. 2009.06, Bank of Israel.

van den Brakel, J. and Krieg, S. (2009). Estimation of the Monthly Unemployment Rate Through Structural Time Series Modelling in a Rotating Panel Design. *Survey Methodology*, 35(2):177–190.

van den Brakel, J. A. and Krieg, S. (2015). Dealing with Small Sample Sizes, Rotation Group Bias and Discontinuities in a Rotating Panel Design. *Survey Methodology*, 41(2):267–296.

van den Brakel, J. A. and Krieg, S. (2016). Small Area Estimation with State Space Common Factor Models for Rotating Panels. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(3):763–791.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.

A State space representations

For the sake of simplicity, in this Appendix the subscript t (without the superscript k) indicates that the model is expressed at the low (monthly) frequency.

A.1 Labour force model with univariate auxiliary series

Throughout this section it is assumed that the univariate auxiliary series are the claimant counts, therefore $x_t = x_t^{CC}$.

The observation equation is:

$$\underset{6\times 1}{\begin{pmatrix} \boldsymbol{y}_t \\ x_t \end{pmatrix}} = \boldsymbol{Z}_t \begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \boldsymbol{\alpha}_t^x \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \varepsilon_t^x \end{pmatrix} = \begin{bmatrix} \boldsymbol{Z}_t^y & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Z}^x \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \boldsymbol{\alpha}_t^x \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \varepsilon_t^x \end{pmatrix}, \quad \begin{pmatrix} \mathbf{0} \\ \varepsilon_t^x \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{H}),$$

$\underset{6\times 6}{\boldsymbol{H}} = \operatorname{diag}\left(\mathbf{0}', \sigma^2_{\varepsilon,x}\right)$.

The state variables for $\boldsymbol{y}_t$ (i.e., the level, the slope, the seasonality, the RGB and the survey errors) are:

$$\underset{30\times 1}{\boldsymbol{\alpha}_t^y} = \big(L_t^y \quad R_t^y \quad S_{1,t}^y \quad S_{1,t}^{*y} \quad S_{2,t}^y \quad S_{2,t}^{*y} \quad S_{3,t}^y \quad S_{3,t}^{*y} \quad S_{4,t}^y \quad S_{4,t}^{*y} \quad S_{5,t}^y \quad S_{5,t}^{*y} \quad S_{6,t}^y \quad \lambda_{2,t} \quad \lambda_{3,t} \quad \lambda_{4,t} \quad \lambda_{5,t} \quad \boldsymbol{\alpha}_{E,t}'\big)'$$

$$\underset{13\times 1}{\boldsymbol{\alpha}_{E,t}} = \begin{pmatrix} \tilde{e}_{1,t} & \tilde{e}_{2,t} & \tilde{e}_{3,t} & \tilde{e}_{4,t} & \tilde{e}_{5,t} & \tilde{e}_{1,t-2} & \tilde{e}_{2,t-2} & \tilde{e}_{3,t-2} & \tilde{e}_{4,t-2} & \tilde{e}_{1,t-1} & \tilde{e}_{2,t-1} & \tilde{e}_{3,t-1} & \tilde{e}_{4,t-1} \end{pmatrix}',$$

where E refers to the structure of the autocorrelated sampling errors that are modelled as state variables.

The state variables for x_t (i.e., the level, the slope and the seasonality) are:

$$\underset{13\times 1}{\boldsymbol{\alpha}_t^x} = \begin{pmatrix} L_t^x & R_t^x & S_{1,t}^x & S_{1,t}^{*x} & S_{2,t}^x & S_{2,t}^{*x} & S_{3,t}^x & S_{3,t}^{*x} & S_{4,t}^x & S_{4,t}^{*x} & S_{5,t}^x & S_{5,t}^{*x} & S_{6,t}^x \end{pmatrix}'.$$

$$\underset{5\times 30}{\boldsymbol{Z}_t^y} = \begin{bmatrix} 1&0&1&0&1&0&1&0&1&0&1&0&1&0&0&0&0 & \\ 1&0&1&0&1&0&1&0&1&0&1&0&1&1&0&0&0 & \\ 1&0&1&0&1&0&1&0&1&0&1&0&1&0&1&0&0 & \boldsymbol{Z}_{E,t}^y \\ 1&0&1&0&1&0&1&0&1&0&1&0&1&0&0&1&0 & \\ 1&0&1&0&1&0&1&0&1&0&1&0&1&0&0&0&1 & \end{bmatrix},$$

$$\underset{5\times 13}{\boldsymbol{Z}_{E,t}^y} = \begin{bmatrix} c_{1,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & c_{2,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & c_{3,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & c_{4,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & c_{5,t} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\underset{1\times 13}{\boldsymbol{Z}^x} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

The transition equation takes the form:

$$\underset{43\times 1}{\begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \boldsymbol{\alpha}_t^x \end{pmatrix}} = \boldsymbol{T} \begin{pmatrix} \boldsymbol{\alpha}_{t-1}^y \\ \boldsymbol{\alpha}_{t-1}^x \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^y \\ \boldsymbol{\eta}_t^x \end{pmatrix} = \begin{bmatrix} \boldsymbol{T}^y & \mathbf{0} \\ \mathbf{0} & \boldsymbol{T}^x \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_{t-1}^y \\ \boldsymbol{\alpha}_{t-1}^x \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^y \\ \boldsymbol{\eta}_t^x \end{pmatrix}.$$

The transition matrix for $\boldsymbol{y}_t$ is:

$$\underset{30\times 30}{\boldsymbol{T}^y} = \operatorname{blockdiag}(\boldsymbol{T}_\mu^y, \boldsymbol{T}_\omega^y, \boldsymbol{T}_\lambda^y, \boldsymbol{T}_E^y).$$

The transition matrix for the level and slope components is:

$$\underset{2\times 2}{\boldsymbol{T}_\mu^y} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

The transition matrix for the seasonal component is:

$$\underset{11\times11}{\boldsymbol{T}^{y}_{\omega}} = \text{blockdiag}(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{C}_4, \boldsymbol{C}_5, -1),$$

$$\boldsymbol{C}_j = \begin{bmatrix} \cos(h_l) & \sin(h_l) \\ -\sin(h_l) & \cos(h_l) \end{bmatrix}, \quad h_l = \pi l/6, \quad l = 1, ..., 6.$$

The transition matrix for the RGB component is:

$$\underset{4\times4}{\boldsymbol{T}^{y}_{\lambda}} = \boldsymbol{I}_4.$$

The transition matrix for the autocorrelated survey errors is:

$$\underset{13\times13}{\boldsymbol{T}^{y}_{E}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \delta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The transition matrix for x_t, $\underset{13\times13}{\boldsymbol{T}^{x}}$, is the same as $\boldsymbol{T}^{y}$ without the transition matrices for the RGB component and for the survey errors.

The vector of innovations is defined as follows:

$$\underset{30\times1}{\boldsymbol{\eta}^{y}_{t}} = \big(\eta^{y}_{L,t} \quad \eta^{y}_{R,t} \quad \eta^{y}_{\omega,1,t} \quad \eta^{*y}_{\omega,1,t} \quad \eta^{y}_{\omega,2,t} \quad \eta^{*y}_{\omega,2,t} \quad \eta^{y}_{\omega,3,t} \quad \eta^{*y}_{\omega,3,t} \quad \eta^{y}_{\omega,4,t} \quad \eta^{*y}_{\omega,4,t}$$
$$\eta^{y}_{\omega,5,t} \quad \eta^{*y}_{\omega,5,t} \quad \eta^{y}_{\omega,6,t} \quad \eta_{\lambda,2,t} \quad \eta_{\lambda,3,t} \quad \eta_{\lambda,4,t} \quad \eta_{\lambda,5,t} \quad \boldsymbol{\eta}'^{y}_{E,t}\big)',$$

$$\underset{13\times1}{\boldsymbol{\eta}^{y}_{E,t}} = \begin{pmatrix} \nu_{1,t} & \nu_{2,t} & \nu_{3,t} & \nu_{4,t} & \nu_{5,t} & \mathbf{0}' \end{pmatrix}',$$

$$\underset{13\times1}{\boldsymbol{\eta}^{x}_{t}} = \begin{pmatrix} \eta^{x}_{L,t} & \eta^{x}_{R,t} & \eta^{x}_{\omega,1,t} & \eta^{*x}_{\omega,1,t} & \eta^{x}_{\omega,2,t} & \eta^{*x}_{\omega,2,t} & \eta^{x}_{\omega,3,t} & \eta^{*x}_{\omega,3,t} & \eta^{x}_{\omega,4,t} & \eta^{*x}_{\omega,4,t} & \eta^{x}_{\omega,5,t} & \eta^{*x}_{\omega,5,t} & \eta^{x}_{\omega,6,t} \end{pmatrix}',$$

$$\underset{43\times1}{\boldsymbol{\eta}_{t}} = \begin{pmatrix} \boldsymbol{\eta}'^{y}_{t} & \boldsymbol{\eta}'^{x}_{t} \end{pmatrix}' \sim N(\mathbf{0}, \boldsymbol{Q}),$$

$$\underset{43\times43}{\boldsymbol{Q}} = \begin{bmatrix} \sigma^2_{L,y} & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & 0 & \mathbf{0}' \\ 0 & \sigma^2_{R,y} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & \rho\sigma_{R,y}\sigma_{R,x} & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q}^{y}_{\omega} & \underset{11\times4}{\mathbf{0}} & \underset{11\times5}{\mathbf{0}} & \underset{11\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{11\times11}{\mathbf{0}} \\ \mathbf{0} & \mathbf{0} & \underset{4\times11}{\mathbf{0}} & \boldsymbol{Q}^{y}_{\lambda} & \underset{4\times5}{\mathbf{0}} & \underset{4\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{4\times11}{\mathbf{0}} \\ \mathbf{0} & \mathbf{0} & \underset{5\times11}{\mathbf{0}} & \underset{5\times4}{\mathbf{0}} & \boldsymbol{Q}^{y}_{\nu} & \underset{5\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{5\times11}{\mathbf{0}} \\ \mathbf{0} & \mathbf{0} & \underset{8\times11}{\mathbf{0}} & \underset{8\times4}{\mathbf{0}} & \underset{8\times5}{\mathbf{0}} & \underset{8\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{8\times11}{\mathbf{0}} \\ 0 & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \sigma^2_{L,x} & 0 & \mathbf{0}' \\ 0 & \rho\sigma_{R,y}\sigma_{R,x} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & \sigma^2_{R,x} & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \underset{11\times11}{\mathbf{0}} & \underset{11\times4}{\mathbf{0}} & \underset{11\times5}{\mathbf{0}} & \underset{11\times8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \boldsymbol{Q}^{x}_{\omega} \end{bmatrix},$$

where $\sigma^2_{L,y} = \sigma^2_{L,x} = 0$ in the Dutch labour force model, $\underset{11\times11}{\boldsymbol{Q}^z_\omega} = \sigma^2_{\omega,z}\boldsymbol{I}_{11}$, for $z = x, y$, $\underset{4\times4}{\boldsymbol{Q}^y_\lambda} = \sigma^2_\lambda \boldsymbol{I}_4$ and $\underset{5\times5}{\boldsymbol{Q}^y_\nu} = \operatorname{diag}\left(\sigma^2_{\nu_1}, \sigma^2_{\nu_2}, \sigma^2_{\nu_3}, \sigma^2_{\nu_4}, \sigma^2_{\nu_5}\right)$.

A.2 Labour force model with high-dimensional auxiliary series

Throughout this section it is assumed that the high-dimensional auxiliary series are the Google Trends, therefore $\boldsymbol{x}_t = \boldsymbol{x}^{GT}_t$. n is the number of Google Trends. It is assumed only $r = 1$ factor for the Google Trends.

The observation equation is:

$$\underset{(5+n)\times1}{\begin{pmatrix} \boldsymbol{y}_t \\ \boldsymbol{x}_t \end{pmatrix}} = \underset{(5+n)\times31}{\boldsymbol{Z}_t} \begin{pmatrix} \boldsymbol{\alpha}^y_t \\ \alpha^x_t \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\varepsilon}_t \end{pmatrix} = \begin{bmatrix} \boldsymbol{Z}^y_t & \mathbf{0} \\ \underset{n\times31}{\mathbf{0}} & \underset{n\times1}{\hat{\Lambda}} \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}^y_t \\ f_t \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\varepsilon}_t \end{pmatrix}, \quad \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\varepsilon}_t \end{pmatrix} \sim N\left(\mathbf{0}, \hat{\boldsymbol{H}}\right),$$

$$\underset{(5+n)\times(5+n)}{\hat{\boldsymbol{H}}} = \operatorname{diag}\left(\mathbf{0}', \operatorname{diag}\left(\underset{n\times n}{\hat{\boldsymbol{\Psi}}}\right)\right).$$

$\boldsymbol{Z}^y_t$ is the same as in Appendix A.1.

The transition equation takes the form:

$$\underset{31\times1}{\begin{pmatrix} \boldsymbol{\alpha}^y_t \\ f_t \end{pmatrix}} = \underset{31\times31}{\boldsymbol{T}} \begin{pmatrix} \boldsymbol{\alpha}^y_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}^y_t \\ \eta^x_t \end{pmatrix} = \begin{bmatrix} \boldsymbol{T}^y & \mathbf{0} \\ \mathbf{0} & \boldsymbol{T}^x \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}^y_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}^y_t \\ u_t \end{pmatrix}.$$

$\boldsymbol{T}^y$ is the same as in Appendix A.1, and $\boldsymbol{T}^x = 1$.

The vector of innovations is:

$$\underset{31\times1}{\boldsymbol{\eta}_t} = \begin{pmatrix} \boldsymbol{\eta}^{\prime y}_t & u_t \end{pmatrix}' \sim N\left(\mathbf{0}, \boldsymbol{Q}\right),$$

$$\underset{31\times31}{\boldsymbol{Q}} = \begin{bmatrix} \sigma^2_{L,y} & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 \\ 0 & \sigma^2_{R,y} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \rho\sigma_{R,y}\sigma_u \\ \mathbf{0} & \mathbf{0} & \underset{11\times11}{\boldsymbol{Q}^y_\omega} & \underset{11\times4}{\mathbf{0}} & \underset{11\times5}{\mathbf{0}} & \underset{11\times8}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{4\times11}{\mathbf{0}} & \underset{4\times4}{\boldsymbol{Q}^y_\lambda} & \underset{4\times5}{\mathbf{0}} & \underset{4\times8}{\mathbf{0}} & \mathbf{0} \\ 0 & 0 & \underset{5\times11}{\mathbf{0}} & \underset{5\times4}{\mathbf{0}} & \underset{5\times5}{\boldsymbol{Q}^y_\nu} & \underset{5\times8}{\mathbf{0}} & 0 \\ \mathbf{0} & \mathbf{0} & \underset{8\times11}{\mathbf{0}} & \underset{8\times4}{\mathbf{0}} & \underset{8\times5}{\mathbf{0}} & \underset{8\times8}{\mathbf{0}} & \mathbf{0} \\ 0 & \rho\sigma_{R,y}\sigma_u & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \sigma^2_u \end{bmatrix},$$

where $\boldsymbol{\eta}^y_t$ and the first (30×30) diagonal elements of $\boldsymbol{Q}$ are the same as in Appendix A.1.

A.2.1 Extension of the model to incorporate the lags of f_t

Consider a regression of $\eta^y_{R,t}$ on the past values of u_t:

$$\begin{pmatrix} R^y_t \\ f_t \end{pmatrix} = \begin{pmatrix} R^y_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \eta^y_{R,t} \\ u_t \end{pmatrix}, \quad u_t \sim N\left(0, \sigma^2_u\right),$$

$$\eta^y_{R,t} = \sum_{j=1}^{q} \kappa_j u_{t-j} + w_t = \kappa_1 f_{t-1} + \sum_{j=2}^{q} \left(\kappa_j - \kappa_{j-1}\right) f_{t-j} - \kappa_q f_{t-q-1} + w_t, \quad w_t \sim N\left(0, \sigma^2_w\right).$$

The transition equation (for simplicity, without the seasonal, RGB and survey error components) becomes:

$$\begin{pmatrix} L_t^y \\ R_t^y \\ f_t \\ f_{t-1} \\ f_{t-2} \\ \vdots \\ f_{t-q} \end{pmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & \mathbf{0}' & 0 & 0 \\ 0 & 1 & \kappa_1 & (\kappa_2-\kappa_1) & \dots & (\kappa_q-\kappa_{q-1}) & -\kappa_q \\ 0 & 0 & 1 & 0 & \mathbf{0}' & 0 & 0 \\ 0 & 0 & 1 & 0 & \mathbf{0}' & 0 & 0 \\ 0 & 0 & 0 & 1 & \mathbf{0}' & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \mathbf{0} \\ 0 & 0 & 0 & 0 & \mathbf{0}' & 1 & 0 \end{bmatrix} \begin{pmatrix} L_{t-1}^y \\ R_{t-1}^y \\ f_{t-1} \\ f_{t-2} \\ f_{t-3} \\ \vdots \\ f_{t-q-1} \end{pmatrix} + \begin{pmatrix} 0 \\ w_t \\ u_t \\ 0 \\ 0 \\ \mathbf{0} \\ 0 \end{pmatrix},$$

$$\begin{pmatrix} w_t \\ u_t \end{pmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_w^2 & \rho\sigma_w\sigma_u \\ \rho\sigma_w\sigma_u & \sigma_u^2 \end{bmatrix}\right).$$

In the measurement equation $\boldsymbol{Z}^x = \underset{n\times 1}{\hat{\Lambda}} \quad \underset{n\times q}{\mathbf{0}}$.

A.2.2 Extension of the model to incorporate the seasonality/cycle in f_t with a (seasonal) ARIMA model

Assume an ARIMA$(3,1,1)$ process for f_t:

$$f_t = f_{t-1} + \phi_1(f_{t-1} - f_{t-2}) + \phi_2(f_{t-2} - f_{t-3}) + \phi_3(f_{t-3} - f_{t-4}) + u_t + \gamma u_{t-1}, \quad u_t \sim N(0,1).$$

The state space representation of the above model is based on Durbin and Koopman (2012) and illustrated below. Let $\boldsymbol{f}_t$ be the state vector

$$\boldsymbol{f}_t = \begin{pmatrix} f_{t-1} \\ f_t - f_{t-1} \\ \phi_2(f_{t-1} - f_{t-2}) \\ \phi_3(f_{t-2} - f_{t-3}) + \gamma u_t \end{pmatrix}.$$

The transition equation for $\boldsymbol{f}_t$ takes the form:

$$\boldsymbol{f}_t = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & \phi_1 & 1 & 1 \\ 0 & \phi_2 & 0 & 0 \\ 0 & 0 & \frac{\phi_3}{\phi_2} & 0 \end{bmatrix} \boldsymbol{f}_{t-1} + \begin{pmatrix} 0 \\ 1 \\ 0 \\ \gamma \end{pmatrix} u_t.$$

Consequently, the observation equation becomes:

$$\boldsymbol{x}_t = \hat{\boldsymbol{\Lambda}} \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix} + \boldsymbol{\varepsilon}_t.$$

Note that the transition equation of the full state space model is now expressed in the form:

$$\boldsymbol{\alpha}_t = \boldsymbol{T}\boldsymbol{\alpha}_{t-1} + \boldsymbol{R}\boldsymbol{\eta}_t,$$

where

$$\underset{\dim(\boldsymbol{\alpha}_t)\times\dim(\boldsymbol{\alpha}_t)}{\boldsymbol{R}} = \begin{bmatrix} \boldsymbol{I}_{\dim(\boldsymbol{\alpha}_t)-4} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & 0 & 0 & 0 & 0 \\ \mathbf{0}' & 0 & 1 & 0 & 0 \\ \mathbf{0}' & 0 & 0 & 0 & 0 \\ \mathbf{0}' & 0 & 0 & 0 & \gamma \end{bmatrix}.$$

We here allow u_t to be correlated with $\eta_{R,t}^y$.

A.2.3 I(1) idiosyncratic components

Consider the following toy example to have a clearer understanding of the estimation procedure when some of the idiosyncratic components are $I(1)$.

$$\boldsymbol{x}_t = \Lambda f_t + \boldsymbol{\varepsilon}_t.$$

Suppose that $\boldsymbol{x}_t$ and $\boldsymbol{\varepsilon}_t$ are 5-dimensional vectors ($n = 5$), and f_t is univariate. Suppose that $\varepsilon_{1,t}$ and $\varepsilon_{3,t}$ are $I(1)$, whereas $\varepsilon_{2,t}$, $\varepsilon_{4,t}$ and $\varepsilon_{5,t}$ are $I(0)$. Then the observation equation for $\boldsymbol{x}_t$ becomes:

$$\begin{pmatrix} x_{1,t} \\ x_{2,t} \\ x_{3,t} \\ x_{4,t} \\ x_{5,t} \end{pmatrix} = \begin{bmatrix} \Lambda_1 & 1 & 0 \\ \Lambda_2 & 0 & 0 \\ \Lambda_3 & 0 & 1 \\ \Lambda_4 & 0 & 0 \\ \Lambda_5 & 0 & 0 \end{bmatrix} \begin{pmatrix} f_t \\ \varepsilon_{1,t} \\ \varepsilon_{3,t} \end{pmatrix} + \begin{pmatrix} 0 \\ \varepsilon_{2,t} \\ 0 \\ \varepsilon_{4,t} \\ \varepsilon_{5,t} \end{pmatrix},$$

where f_t, $\varepsilon_{1,t}$ and $\varepsilon_{3,t}$ are state variables with transition equation

$$\begin{pmatrix} f_t \\ \varepsilon_{1,t} \\ \varepsilon_{3,t} \end{pmatrix} = \boldsymbol{I}_3 \begin{pmatrix} f_{t-1} \\ \varepsilon_{1,t-1} \\ \varepsilon_{3,t-1} \end{pmatrix} + \begin{pmatrix} u_t \\ \xi_{1,t} \\ \xi_{3,t} \end{pmatrix}.$$

$$\boldsymbol{\Psi} = \operatorname{cov}\begin{pmatrix} \xi_{1,t} & \varepsilon_{2,t} & \xi_{3,t} & \varepsilon_{4,t} & \varepsilon_{5,t} \end{pmatrix}' = \operatorname{cov}\begin{pmatrix} \Delta\varepsilon_{1,t} & \varepsilon_{2,t} & \Delta\varepsilon_{3,t} & \varepsilon_{4,t} & \varepsilon_{5,t} \end{pmatrix}'.$$

The covariance matrix between the innovation terms in the observation equation is

$$\operatorname{cov}\begin{pmatrix} 0 & \varepsilon_{2,t} & 0 & \varepsilon_{4,t} & \varepsilon_{5,t} \end{pmatrix}' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \psi_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \psi_{44} & 0 \\ 0 & 0 & 0 & 0 & \psi_{55} \end{pmatrix},$$

and ends up in the $\boldsymbol{H}$ matrix defined in Appendices A.2 or A.3. On the contrary, the covariance matrix between the innovations of the state variables is

$$\operatorname{cov}\begin{pmatrix} u_t & \xi_{1,t} & \xi_{3,t} \end{pmatrix}' = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \psi_{11} & 0 \\ 0 & 0 & \psi_{33} \end{pmatrix},$$

and ends up in the $\boldsymbol{Q}$ matrix defined in Appendices A.2 or A.3.

A.3 Labour force model with univariate and high-dimensional auxiliary series

Throughout this section both the claimant counts and the Google Trends are included in the model as auxiliary series.

The observation equation is:

$$\underset{(6+n)\times 1}{\begin{pmatrix} \boldsymbol{y}_t \\ x_t^{CC} \\ \boldsymbol{x}_t^{GT} \end{pmatrix}} = \underset{(6+n)\times 44}{\boldsymbol{Z}_t} \begin{pmatrix} \boldsymbol{\alpha}_t^{y} \\ \boldsymbol{\alpha}_t^{CC} \\ \alpha_t^{GT} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \varepsilon_t^{CC} \\ \boldsymbol{\varepsilon}_t \end{pmatrix} = \begin{bmatrix} \underset{5\times 13}{\boldsymbol{Z}_t^{y}} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0}' & \underset{1\times 13}{\boldsymbol{Z}^{CC}} & 0 \\ \underset{n\times 30}{\boldsymbol{0}} & \underset{n\times 13}{\boldsymbol{0}} & \underset{n\times 1}{\hat{\Lambda}} \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_t^{y} \\ \boldsymbol{\alpha}_t^{CC} \\ f_t \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} \\ \varepsilon_t^{CC} \\ \boldsymbol{\varepsilon}_t^{GT} \end{pmatrix},$$

$$\begin{pmatrix} \boldsymbol{0} \\ \varepsilon_t^{CC} \\ \boldsymbol{\varepsilon}_t^{GT} \end{pmatrix} \sim N\left(\boldsymbol{0}, \boldsymbol{H}\right), \qquad \underset{(6+n)\times(6+n)}{\boldsymbol{H}} = \operatorname{diag}\left(\boldsymbol{0}', \sigma^2_{\varepsilon,x}, \operatorname{diag}\left(\underset{n\times n}{\hat{\boldsymbol{\Psi}}}\right)\right).$$

$\boldsymbol{Z}_t^y$ is the same as in Appendix A.1, and $\boldsymbol{Z}^{CC}$ is the same as $\boldsymbol{Z}^x$ in Appendix A.1.

The transition equation takes the form:

$$\underset{44\times 1}{\begin{pmatrix} \boldsymbol{\alpha}_t^y \\ \boldsymbol{\alpha}_t^{CC} \\ f_t \end{pmatrix}} = \underset{44\times 44}{\boldsymbol{T}} \begin{pmatrix} \boldsymbol{\alpha}_{t-1}^y \\ \boldsymbol{\alpha}_{t-1}^{CC} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^y \\ \boldsymbol{\eta}_t^{CC} \\ \eta_t^{GT} \end{pmatrix} = \begin{bmatrix} \boldsymbol{T}^y & \underset{30\times 13}{\mathbf{0}} & \mathbf{0} \\ \underset{13\times 30}{\mathbf{0}} & \boldsymbol{T}^{CC} & \mathbf{0} \\ \mathbf{0}' & \mathbf{0}' & 1 \end{bmatrix} \begin{pmatrix} \boldsymbol{\alpha}_{t-1}^y \\ \boldsymbol{\alpha}_{t-1}^{CC} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\eta}_t^y \\ \boldsymbol{\eta}_t^{CC} \\ u_t \end{pmatrix}.$$

$\boldsymbol{T}^y$ is the same as in Appendix A.1, and $\boldsymbol{T}^{CC}$ and $\boldsymbol{\alpha}_t^{CC}$ are, respectively, the same as $\boldsymbol{T}^x$ and $\boldsymbol{\alpha}_t^x$ in Appendix A.1.

The vector of innovations is:

$$\underset{44\times 1}{\boldsymbol{\eta}_t} = \begin{pmatrix} \boldsymbol{\eta}_t^{\prime y} & \boldsymbol{\eta}_t^{\prime CC} & u_t \end{pmatrix}' \sim N(\mathbf{0}, \boldsymbol{Q}),$$

$$\underset{44\times 44}{\boldsymbol{Q}} = \begin{bmatrix} \sigma_{L,y}^2 & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & 0 & \mathbf{0}' & 0 \\ 0 & \sigma_{R,y}^2 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \mathbf{0}' & \rho_{GT}\sigma_{R,y}\sigma_u \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_\omega^y & \underset{11\times 4}{\mathbf{0}} & \underset{11\times 5}{\mathbf{0}} & \underset{11\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{11\times 11}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{4\times 11}{\mathbf{0}} & \boldsymbol{Q}_\lambda^y & \underset{4\times 5}{\mathbf{0}} & \underset{4\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{4\times 11}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{5\times 11}{\mathbf{0}} & \underset{5\times 4}{\mathbf{0}} & \boldsymbol{Q}_\nu^y & \underset{5\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{5\times 11}{\mathbf{0}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \underset{8\times 11}{\mathbf{0}} & \underset{8\times 4}{\mathbf{0}} & \underset{8\times 5}{\mathbf{0}} & \underset{8\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \underset{8\times 11}{\mathbf{0}} & \mathbf{0} \\ 0 & 0 & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \sigma_{L,CC}^2 & 0 & \mathbf{0}' & 0 \\ 0 & \rho_{CC}\sigma_{R,y}\sigma_{R,CC} & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & \sigma_{R,CC}^2 & \mathbf{0}' & 0 \\ \mathbf{0} & \mathbf{0} & \underset{11\times 11}{\mathbf{0}} & \underset{11\times 4}{\mathbf{0}} & \underset{11\times 5}{\mathbf{0}} & \underset{11\times 8}{\mathbf{0}} & \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_\omega^{CC} & \mathbf{0} \\ 0 & \rho_{GT}\sigma_{R,y}\sigma_u & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & 0 & 0 & \mathbf{0}' & \sigma_u^2 \end{bmatrix},$$

where $\boldsymbol{\eta}_t^y$ is the same as in Appendix A.1. $\boldsymbol{\eta}_t^{CC}$ and $\sigma_{R,CC}$ are respectively the same as $\boldsymbol{\eta}_t^x$ and $\sigma_{R,x}$ in Appendix A.1. The first (43×43) elements of $\boldsymbol{Q}$ are the same as in Appendix A.1, whereas the last row and column are the same as in Appendix A.2.